*Astronomy & Astrophysics* manuscript no. manuscript
May 31, 2018


# Measuring precise radial velocities and cross-correlation function line-profile variations using a Skew Normal density ⋆

U. Simola[1,2] ⋆⋆, X. Dumusque[3] ⋆⋆⋆, and Jessi Cisewski-Kehe[2]

[1] Department of Statistical Sciences, University of Padua, Padua, Italy
[2] Department of Statistics and Data Science, Yale University, New Haven, CT, USA
[3] Observatoire de Genève, Université de Genève, 51 ch. des Maillettes, CH-1290 Versoix, Switzerland

Received XXX; accepted XXX

**ABSTRACT**

*Context.* Stellar activity is one of the primary limitations to the detection of low-mass exoplanets using the radial-velocity (RV) technique. Stellar activity can be probed by measuring time dependent variations in the shape of the cross-correlation function (CCF), often estimated using the different moments of the modeled CCF. Therefore estimating the moments of the CCF with high precision is essential to de-correlate exoplanet signals from spurious RV signals originating from stellar activity.
*Aims.* We propose to estimate the moments of the CCF by fitting a model using a Skew Normal (SN) density shape, which unlike the commonly employed Normal density, includes an skewness parameter to capture the asymmetry of the CCF induced by stellar activity, but also the natural asymmetry induced by convective blueshift.
*Methods.* The performance of the proposed method is compared to the Normal density using both simulated and real observations with varying levels of activity and signal-to-noise ratio (SNR) levels.
*Results.* When considering the real observations, the correlation between the RV's and the asymmetry of the CCF and the correlation between the RV's and the width of the CCF are stronger when using the parameters derived from the SN than the Normal approach. This suggests that the CCF asymmetry and the CCF width derived using a SN may be more sensitive to stellar activity, which can be helpful with estimating stellar rotational periods and generally characterizing the stellar activity signals. The estimated uncertainties in the estimated RV's using the proposed SN approach are on average 10% smaller than the uncertainties calculated on the mean of the Normal, and the estimated uncertainties on the SN asymmetry parameter are on average 15% smaller than the commonly used Bisector Inverse Slope Span (BIS SPAN) approach for estimating the asymmetry of the CCF.
*Conclusions.* XXX [[**Jessi:** It looks like we still need to add conclusions here]]

**Key words.** techniques: radial velocities – planetary systems – stars: activity – methods: data analysis

## 1. Introduction

When working with radial velocities (RV's), one of the main limitations to the detection of small-mass exoplanets is no longer the precision of the instruments used, but the different sources of variability induced by the stars (e.g. Feng et al. 2017; Dumusque et al. 2017; Rajpaul et al. 2015; Robertson et al. 2014). Stellar oscillations, granulation phenomena, and stellar activity can all induce apparent RV signals (e.g. Saar & Donahue 1997; Queloz et al. 2001; Desort et al. 2007; Dumusque et al. 2011; Dumusque 2016) that are above the meter-per-second precision reached by the best high-resolution spectrographs (HARPS, HARPS-N, Mayor et al. 2003; Cosentino et al. 2012). It is therefore mandatory to better understand stellar signals and to develop methods to correct for them, if in the near future we want to detect or confirm an Earth-twin planet using the RV technique. This is even more true now that instrument like the Echelle SPectrograph for Rocky Exoplanet and Stable Spectroscopic Observations (ESPRESSO) (Pepe et al. 2014) and EXtreme PREcision Spectrometer (EXPRES) (Fischer et al. 2016) should have the stability to detect such signals. However, if solutions are not found to mitigate the impact of stellar activity, the detection or confirmation of potential Earth-twins will be extremely challenging and false detections could plague the field.

One of the most challenging stellar signal to characterize and to correct for is the signal induced by stellar activity. Stellar activity is responsible for creating magnetic regions on the surface of stars, and those regions change locally the temperature and the convection, which can induce spurious RV's variations (Meunier et al. 2010; Dumusque et al. 2014). In theory, it should be easy to differentiate between the pure Doppler-shift induced by a planet, which shifts the entire stellar spectrum, and stellar activity, which modifies the shape of spectral lines and by doing so create a spurious shift of the stellar spectrum (Saar & Donahue 1997; Hatzes 2002; Kurster et al. 2003; Lindegren & Dravins 2003; Desort et al. 2007; Lagrange et al. 2010; Meunier et al. 2010; Dumusque et al. 2014). However, on quiet GKM dwarfs, the main target for precise RV's measurements, stellar activity can induce signals of a few $\mathrm{m\,s^{-1}}$. This corresponds physically to variations smaller than 1/100th of a pixel on the detector making the changing shape of the spectral lines challenging to detect.

In order to measure such tiny variations, a common approach is to average the information of all the lines in the spectrum by cross correlating the stellar spectrum with a synthetic (Baranne et al. 1996; Pepe et al. 2002) or an observed stellar template

⋆ Based on observations collected at the La Silla Parana Observatory, ESO (Chile), with the HARPS spectrograph at the 3.6-m telescope.
⋆⋆ e-mail: umberto.simola@helsinki.fi
⋆⋆⋆ Branco Weiss Fellow–Society in Science (url: http://www.society-in-science.org)

(Anglada-Escudé & Butler 2012). The result of this operation gives us the cross-correlation function (CCF). The CCF gives the spectrum's cross-correlation with the template as the template is shifted according to different RVs. To measure the Doppler-shift between different spectra and therefore to retrieve the RV's of a star as a function of time, the variations of the CCF barycenter are calculated. The barycenter is generally estimated by fitting a Normal density to the CCF and retaining its mean. Variations in line shape between different spectra, which indicate the presence of signals induced by stellar activity, are measured by analyzing the different moments of the CCF. Usually, the width of the CCF is estimated using the full-width half-maximum (FWHM) of the fitted Normal density, and its asymmetry using the the bisector inverse slope span (BIS SPAN, Queloz et al. 2001).

If an apparent RV signal is induced by activity, generally a strong correlation will be observed between the RV and chromospheric activity indicators like $\log(R'_{HK})$ or H-$\alpha$ (Boisse et al. 2009; Dumusque et al. 2012; Robertson et al. 2014), but also between the RV and the FWHM of the CCF or its BIS SPAN (Queloz et al. 2001; Boisse et al. 2009; Queloz et al. 2009; Dumusque 2016). It is therefore common now, that when fitting a Keplerian signal to a set of RVs to look for a planet, the model includes in addition linear dependancies with the $\log(R'_{HK})$, the FWHM and the BIS SPAN (Dumusque et al. 2017; Feng et al. 2017). It is also common to add a Gaussian process to the model to account for the correlated noise induced by stellar activity. The hyperparameters of the Gaussian process can be trained on different activity indicators (Haywood et al. 2014; Rajpaul et al. 2015). It is therefore essential for mitigating stellar activity to obtain activity indicators that are the most correlated with the RV's but also for which we can obtain the best precision.

Several indicators have been developed that are more sensitive to line asymmetry than the BIS SPAN. In Boisse et al. (2011), the authors develop $V_{span}$, which is the difference between the RV measured respectively by fitting a Normal density to the upper and the bottom part of the CCF. This CCF asymmetry parameter is shown to be more sensitive than the BIS SPAN at low signal-to-noise ratio (SNR). Figueira et al. (2013) studied the use of two new indicators, bi-Gauss and $V_{asy}$. The authors were able to show that when using bi-Gauss, the amplitude in asymmetry is 30% larger than when using BIS SPAN, therefore allowing the detection of lower levels of activity. They also demonstrated that $V_{asy}$ seems to be a better indicator of line asymmetry at high SNR, as its correlation with RV is more significant than any of the previously proposed asymmetry indictors.

In all the methods described above, except bi-Gauss, the RV and the FWHM are derived using a Normal density fitted to the CCF, and the asymmetry is estimated using another approach. In this paper we propose to use a Skew Normal (SN) density to estimate with a single fit of the CCF, the RV, the FWHM and the asymmetry of the CCF, as this function includes a skewness parameter (Azzalini 1985).

The paper is organized as follow. In Sec. 2 we introduce the SN density, describe its applicability for modeling the CCF, and study how the SN parameters relate to the RV, FWHM and BIS SPAN of the CCF. In Sec. 3 we propose an expanded linear model used to correct the estimated RV's from the inferred stellar activity, which extends the linear models previously proposed for this purpose (Dumusque et al. 2017; Feng et al. 2017). In Sec. 4 the performance of the SN fit to the CCF is investigated using simulations coming from the Spot Oscillation And Planet (SOAP) 2.0 (Dumusque et al. 2014), followed by an analysis of real observations in Sec. 5. Sec. 6 considers derived error bars for the different estimated CCF parameters, and finally a discussion of the results and conclusions are included in Sec. 7 and Sec. 8, respectively.

## 2. The Skew Normal distribution

The Skew Normal (SN) distribution is a class of probability distributions which includes the Normal distribution as a special case (Azzalini 1985). The SN distribution has, in addition to a location and a scale parameter analogous to the Normal distribution's mean and standard deviation, a third parameter which describes the skewness (i.e. the asymmetry) of the distribution. Considering a random variable $Y \in \mathbb{R}$ (where $\mathbb{R}$ is the real line) which follows a SN distribution with location parameter $\xi \in \mathbb{R}$, scale parameter $\omega \in \mathbb{R}^+$ (i.e., the positive real line), and skewness parameter $\alpha \in \mathbb{R}$, its density at some value $y \in Y$ can be written as

$$SN(y;\xi,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\frac{\alpha(y-\xi)}{\omega}\right), \quad (1)$$

where $\phi$ and $\Phi$ are respectively the density function and the distribution function of a standard Normal distribution[1]. The skewness parameter $\alpha$ quantifies the asymmetry of the SN. Examples of SN densities under different skewness parameter values and the same location and scale parameters ($\xi = 0$ and $\omega = 1$) are displayed in Fig. 1. A usual Normal distribution is the special case of the SN distribution when the skewness parameter $\alpha$ is equal to zero[2]. For reasons related to the interpretation of the parameters in Eq. 1 and computational issues with estimating $\alpha$ near 0, a different parametrization is used in this work, which is referred to as the *centered parametrization* (CP). This CP is much closer to the parametrization of a Normal distribution, as

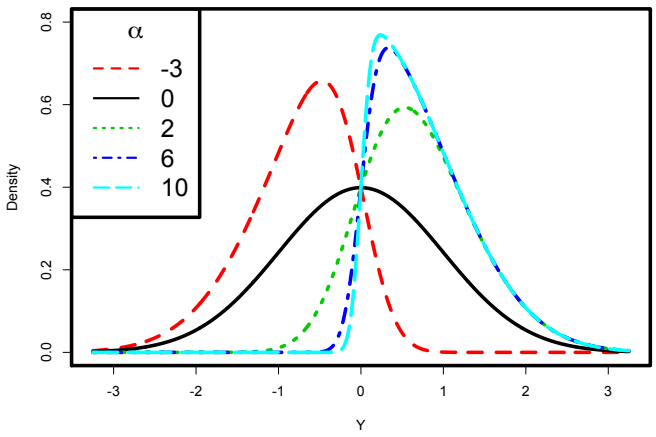


**Fig. 1.** Density function of a random variable Y following the SN distribution $SN(\xi,\omega^2,\alpha)$ with location parameter $\xi = 0$, scale parameter $\omega = 1$ and different values of the skewness parameter $\alpha$ indicated by different colors and line types. Note that the solid black line has an $\alpha = 0$, making it a Normal distribution.

[1] A standard Normal distribution is a Normal distribution with a mean of 0 and a standard deviation of 1.

[2] This can be seen from Eq. 1. If $\alpha = 0$ then $\Phi\left(\frac{\alpha(y-\xi)}{\omega}\right) = \Phi(0) = 0.5$ and therefore $SN(y;\xi,\omega,0) = \frac{1}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)$ which is the density of a Normal distribution. Note that $\Phi(0) = 0.5$ because $\Phi(0)$ is the the probability that a standard Normal random variable is less than or equal than 0.

it uses a mean parameter $\mu$, a variance parameter $\sigma^2$ and a skewness parameter $\gamma$. In order to define the CP, we need to express the CP parameters $(\mu, \sigma^2, \gamma)$ as a function of $(\xi, \omega^2, \alpha)$. This can be done using the following relations:

$$\mu = \xi + \omega\beta, \quad \sigma^2 = \omega^2(1-\beta^2), \quad \gamma = \frac{1}{2}(4-\pi)\beta^3\left(1-\beta^2\right)^{-3/2}, \quad (2)$$

where $\beta = \sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{\sqrt{1+\alpha^2}}\right)$ (e.g. Arellano & Azzalini 2010).

By using Eq. 2, the new set of parameters $(\mu, \sigma^2, \gamma)$ provides a clearer interpretation of the behavior of the SN distribution. For the $\alpha$ values used in Fig. 1, the corresponding values of ($\mu$, $\sigma^2$, $\gamma$) are displayed in Table 1. In particular, $\mu$ and $\sigma^2$ are the actual mean and variance of the distribution, rather than simply a location and scale parameter, and $\gamma$ provides an measure of the skewness of the SN. Along with the mean of the SN, we consider the median of the distribution as a measure of center, which is used in the proposed method. See Table 1 for the medians of the SN densities displayed in Fig. 1.

| $\alpha$ | $\mu$ | $\sigma^2$ | $\gamma$ | Median |
|---|---|---|---|---|
| -3 | -0.757 | 0.427 | -0.667 | -0.672 |
| 0 | 0.000 | 1.000 | 0.000 | 0.000 |
| 2 | 0.714 | 0.491 | 0.454 | 0.655 |
| 6 | 0.787 | 0.381 | 0.891 | 0.674 |
| 10 | 0.794 | 0.370 | 0.956 | 0.674 |

**Table 1.** CP values $(\mu, \sigma^2, \gamma)$ along with the median corresponding to the $\alpha$ values shown in Fig. 1, with location parameter $\xi = 0$ and scale parameter $\omega = 1$. Values are rounded to three decimal places.

Further details about the parametrization from Eq. 1, called the *Direct Parametrization* or DP, the CP, and general statistical properties of the SN are treated in rigorous mathematical and statistical viewpoints in the book by Azzalini & Capitanio (2014).

### 2.1. Fitting the Skew Normal density to the CCF

To fit the CCF using a SN density shape, we use a least-squares algorithm and the following model:

$$f_{CCF}(x_i) = \mathrm{C} - \mathrm{A} \times SN(x_i; \mu, \sigma^2, \gamma), \quad i = 1, \ldots, n \quad (3)$$

where C is an unknown offset for the continuum of the CCF, A is the unknown amplitude of the CCF, commonly referred to as the contrast, and $\mu$, $\sigma^2$ and $\gamma$ are the mean, variance and skewness of the SN as defined above. The values $x_1, \ldots, x_n$ are the different values of the x-axis of the CCF, generally in velocity units (e.g. $\mathrm{m\,s^{-1}}$).

When fitting a Normal density to the CCF, the estimated mean of the model is used as the estimated RV, the FWHM of the Normal density[3] represents the width of the CCF. Because the Normal density is symmetric, the skewness is always equal to 0 so a separate approach is needed to estimate the skewness of the CCF. An estimated skewness parameter is generally obtained by calculating the BIS SPAN of the CCF (see Sect. 1, and e.g. Queloz et al. 2001).

With the proposed SN approach, we propose two estimators of the RV: the mean and median of the SN model fit (referred to as SN mean RV and SN median RV, respectively), and present advantages and limitations for both of these choices in Sec. 5 and Sec. 6. The width of the SN, SN FWHM, is defined in the same way as for the Normal density[4], and finally the skewness of the CCF is estimated by the $\gamma$ parameter.

To evaluate the strength of the correlation between the estimated RV's and the different stellar activity indicators, we calculated the Pearson correlation coefficient, $R$, which in its general form is defined as:

$$R(x, y) = \frac{cov(x, y)}{\sigma(x)\sigma(y)}, \quad (4)$$

where $x$ and $y$ are two quantitative variables, $cov(x, y)$ indicates the covariance between $x$ and $y$, and $\sigma(x)$ and $\sigma(y)$ represent their standard deviations. A $p-$value for the statistical test having null hypothesis $H_0 : R = 0$ is provided, along with a 95% confidence interval for $R$ when needed.

## 3. Radial Velocity correction for stellar activity

Exoplanets only produce a pure RV signal. On the contrary, stellar activity, in particular the presence of active regions on the stellar photosphere, do not produce blueshifts or redshifts of the entire stellar spectrum but can create spurious RV signals by modifying the shape of spectral lines. To track these variations in the shape of the spectral lines, the general approach consists in using the FWHM, the BIS SPAN or other indicators such as those introduced in Boisse et al. (2011) or Figueira et al. (2013), which provide information on the width and asymmetry of the CCF. A strong correlation between the estimated RVs and one or more of these parameters provides an indication that stellar activity signals may be affecting the measurements.

When fitting for planetary signals in RV data, it is common to include linear dependencies with the BIS SPAN and the FWHM and to take into account the signal induced by stellar activity (e.g. Dumusque et al. 2017; Feng et al. 2017). We propose to add additional parameters in the model to correct for stellar activity: first the amplitude parameter A of the CCF, generally referred to as the CCF contrast, and the interaction between the BIS SPAN and the FWHM (or $\gamma$ and SN FWHM in the SN case). The stellar activity correction we propose can therefore be written as:

$$RV_{\text{activity}} = \beta_0 + \beta_1 A + \beta_2 \gamma + \beta_3 \text{SN FWHM} + \beta_4(\gamma \text{SN FWHM}) + \epsilon, \quad (5)$$

where $\beta_0$ is the intercept and $\epsilon$ is the error with mean equal to 0 and covariance matrix equal to $\sigma^2 I$ ($I$ defined as the identity matrix). The contrast parameter $A$ accounts for the presence of a spot on the stellar surface, which produces a change in the amplitude of the CCF and not only on its asymmetry or width (see e.g. Fig. 2 in Dumusque et al. 2014). The benefits of including a variable that quantifies the interaction between $\gamma$ and SN FWHM (or BIS SPAN and FWHM) will be better understood through the results of the examples presented in Sec.4. This interaction term can account for possible interactions between SN FWHM (or FWHM) and $\gamma$ (or BIS SPAN), meaning that each variables' association with the response, $RV_{\text{activity}}$, depends also on the other variable.

The proposed model is analyzed using statistical tests on the parameters $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$ where the null hypothesis is $H_0 : \beta_i = 0$, for $i = 0, \ldots, 4$. The significance level for the tests are set at 0.05. The coefficient of determination, $R^2$, is used

[3] FWHM= $2\sqrt{2\ln 2}\,\sigma$ with standard deviation $\sigma$

[4] Note that SN FWHM does not correspond to the width of the SN density at half maximum like in the Normal case.

to assess how well the proposed linear combination of variables accounts for the variability of $RV_{\text{activity}}$.

[[**Jessi:** Is the following paragraph needed? Is any model selection actually carried out here? It seems the model of Eq. 5 is fit, and then the significance of the variables are discussed, but no attempts are really made to remove/add/step through the various model options. The point about multicollinearity is probably good to keep though.]] When working with a linear regression, there are several ways to select the variables to include in the model. While usually the stepwise technique is used (Efroymson 1960; Hocking 1976), the proposed function defined in Eq. 5, that accounts for stellar activity, is the result of statistical and astronomical considerations. In particular we checked that the correlations between the proposed parameters were not approaching one: if it was the case, the matrix needed to calculate the estimates would be singular, hence non invertible. This problem is known in statistics with the term multicollinearity. A detailed discussion of the topic can be found in the book by Belsley (1991). It is common to see some correlation between the amplitude parameter $A$ and the FWHM (or SN FWHM) of the CCF. However, in the analysis of real data presented in this work, we never observed a correlation coefficient exceeding 0.66 and therefore, the problem of multicollinearity is avoided. Finally, we investigate as well the statistical significant of the interaction term between $A$ and the width, and $A$ and the asymmetry of the CC, however, those interaction were relevant for accounting for stellar signal.

## 4. Simulation Study

In order to evaluate the performance of the proposed SN approach for modelling the CCF and the benefit of using the proposed correction for stellar activity (See Eq. 5), we begin by considering a simulation study using spectra generated from the Spot Oscillation And Planet 2.0 code (SOAP 2.0, Dumusque et al. 2014).

For a given configuration of spots and faculae on the stellar surface, SOAP 2.0 gives as output the simulated CCF as a function of rotational phase. The code also returns the RV and the FWHM by fitting a Normal density to the CCF, and the BIS SPAN by calculating the bisector of the CCF. SOAP 2.0 gives noiseless CCFs affected by stellar activity, which are used to compare the benefits of a SN density fit to the CCF compared to a Normal density fit.

For the simulations discussed below, a star similar to the Sun was modeled. The stellar rotational period is set to 25.0 days, the radius to a solar radius and the star is seen equator on. [[**Jessi:** The previous sentence seems incomplete and I'm not sure how to update it.]] The stellar effective temperature is set to 5778 K (NASA Planetary Fact Sheets)[[**Jessi:** Is there a full citation available for this NASA fact sheet? Also, why is it referenced? Is the effective temperature supposed to be similar to the Sun?]], and a quadratic limb-darkening relation with linear and quadratic coefficients 0.29 and 0.34, respectively (Oshagh et al. 2013; Claret & Bloemen 2011) are used. In order to make the result of the simulations more comparable to real data obtained with the HARPS spectrograph discussed in Sect. 5, the SOAP 2.0 CCFs were generate with a width of 40 $\text{km s}^{-1}$and a resolution of 115'000. [[**Jessi:** Is a resolution of "115'000" standard notation? I'm not sure what it means.]]

### 4.1. Faculae

To see the impact of a facula on the different parameters of the CCF, we simulated the effect of an equatorial faculae [[**Jessi:** Should a footnote be added here to explain that the SOAP 2.0 faculae are not simulated from actual faculae templates?]] of size 3% relative to the visible stellar hemisphere. The faculae is face on when the phase equals to 0. Note that a 3% faculae is relatively large for the Sun; at maximum activity, big faculae have generally a size of 1%. [[**Jessi:** Is there a reference for this that could be added?]] In Fig. 2, we compare the barycentric variation of the CCF as measured when fitting a Normal density and using its mean (RV [[**Jessi:** I believe this was going to be updated to "N mean RV"?]]), and when fitting a SN density and taking its mean (SN mean RV) or its median (SN median RV). We see that all the different estimates of the CCF barycenter present a signal of similar amplitude, however the signal obtained with SN mean RV is different from the two others with a maximum amplitude happening at a different phase.

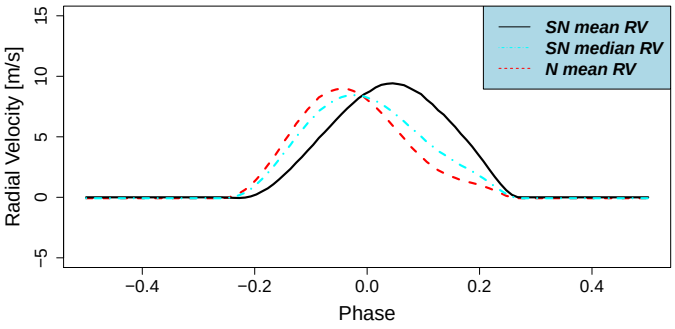


**Fig. 2.** Using CCFs from SOAP 2.0 with a single facula, the estimated RV changes as function of the orbital phase. SN mean RV seems to have the smallest spurious variations caused by the faculae.[[**Jessi:** The previous sentence does not seem correct.]] [[**Jessi:** The cyan line is hard to see in the legend, and may be impossible in a black-and-white version - could the legend background be white?]]

Correlations between the different RV estimates and the different CCF asymmetry or width estimates are displayed in Fig. 3. The strength of the correlation between $\gamma$ and SN mean RV, and $\gamma$ and SN median RV are stronger than the correlations between BIS SPAN and RV, with Pearson correlation coefficient $R$ values of 0.46, -0.67 and -0.09, respectively. For the width barycenter correlations, there is a stronger correlation between SN FWHM and SN mean RV compared to the one between FWHM and RV [[**Jessi:** updated to N mean RV?]], $R = 0.98$ and 0.84, respectively. In this case however, the correlation between SN FWHM and SN median RV is smaller with $R = 0.50$. This first analysis shows that in the case of a facula, using some parameters from the SN can lead to much stronger correlation than the usual Normal parameters and therefore, the SN parameters may better probe stellar activity. We investigate this feature further in the next sections where we consider simulated data with a single spot and a spot plus a planet, and in Sec 5 with real observations.

Since the RV variation displayed in Fig. 2 is caused by only stellar activity, in this case a facula, we applied the activity correction proposed in Eq. 5 to check its performance in this setting. The results of this correction are displayed in Fig. 4 and the statistical tests on the coefficients involved in Eq. 5 are summarized in Table 2. The proposed correction for stellar activity is able to account for the majority of the activity signal created

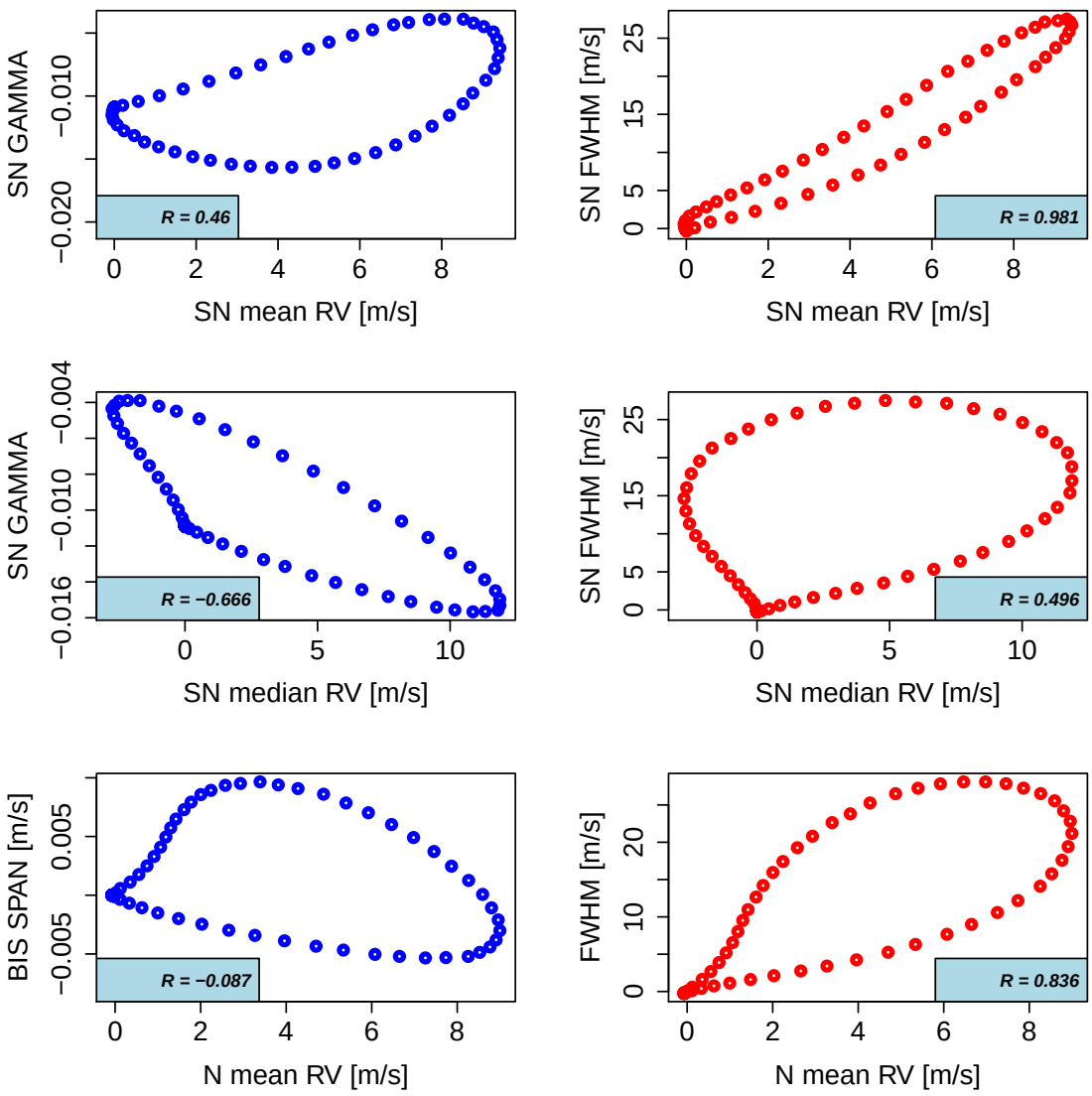


**Fig. 3.** Using CCFs from SOAP 2.0 with a single facula, the correlation between the estimated RV's and the asymmetry parameters are displayed. In this case both the shape and the width of the CCF changes as the facula moves, producing statistically significant correlations between the estimated RV's and, respectively, the asymmetry parameter or the width parameter.

by a facula, with a $R^2$ of our model larger than 0.95. In addition, the rms of the different estimates of the RV reduces from about 3 $\mathrm{m\,s^{-1}}$before correction to values below 0.15 $\mathrm{m\,s^{-1}}$after correction. We see a slightly smaller rms after correction for the SN parameters, however the different is not significant. [[**Jessi:** In the previous sentence, is this a comparison between SN and N? If so, it should be clarified.]] When comparing the correction proposed in Eq. 5 with what is generally used (i.e. a linear combination of only the asymmetry and width parameter), we see that the proposed correction is able to reduce the rms of the RV residuals by a factor of 2. Looking at the significance of the coefficients in table 2, we observe that the parameter related to the intercept, $\beta_0$, is only significant at a level of 1% in the case of the Normal parameters or when SN median RV is used. [[**Jessi:** Is the point about $\beta_0$ relevant?]]

| Parameter | N mean RV | SN mean RV | SN median RV |
|---|---|---|---|
| $\beta_0$ | 0.033 | 0.00020 | 0.61 |
| $\beta_1$ | $2.22e-16$ | $2.22e-16$ | $2.22e-16$ |
| $\beta_2$ | 0.0034 | $2.22e-16$ | $2.22e-16$ |
| $\beta_3$ | 0.00016 | $1.091e-6$ | $9.75e-7$ |
| $\beta_4$ | $2.22e-16$ | $2.22e-16$ | $2.22e-16$ |
| $R^2$ | 0.9978 | 0.9985 | 0.9981 |

**Table 2.** The p-values and coefficient of determination for the models fit using Eq. 5 to correct the estimated RVs from spurious variations caused by a facula. All the variables, except the intercept when the response variable is SN median RV, are statistically significant to explain spurious variations in RV's caused by a facula. The estimated $R^2$ show that the proposed correction for stellar activity explains the vast majority of the spurious variability in the estimated RVs.

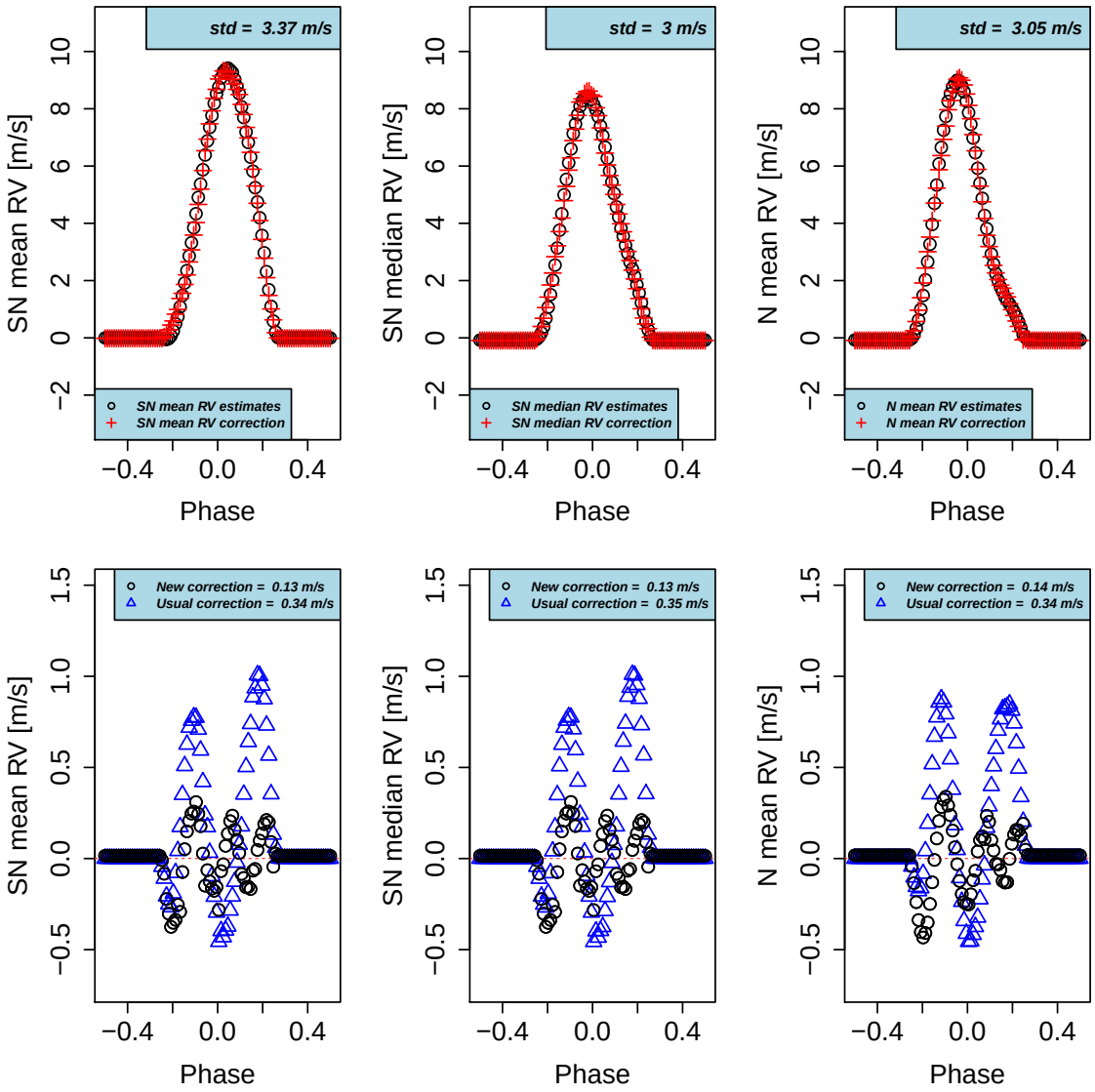


**Fig. 4.** (top) The spurious RVs caused by a facula in the simulated data using a Normal and a SN fit. before and once corrected from stellar activity. [[**Jessi:** What is the correction done here? I think what you mean is this: "The black circles are the estimated RVs and the red plus signs are the estimated RVs from the fit model of Eq. 5." If my assumption is correct, the legend should say "RV estimates" and "RV model" or something like that. Otherwise this is rather confusing because in the bottom plot you are actually comparing residuals from two models, while the top has a different interpretation.]] (bottom) The residuals from the model fit using Eq. 5 (new correction) and the usual correction [[**Jessi:** it might be helpful to actually have an equation for the usual/old model]]. The residuals have a smaller systematic component when using the proposed model of Eq. 5 (blue circles) compared to the usual model (blue triangles). The estimated parameters are presented in Table 2 [[**Jessi:** This table only has p-values and $R^2$]].

### *4.2. Spot*

In this section, we consider the effects on the CCF parameters of an equatorial spot of size 1% relative to the visible stellar hemisphere. The spot is face on when the phase equals to 0. Note that this is a large spot for the Sun, as in general large spots are more in the regime of 0.1%.[[**Jessi:** Can we add a reference for the previous statement?]] In Fig. 5, we shows the barycentric variation of the CCF induced by this simulated spot. In contrast to the case of the facula, all the different estimates of the CCF barycenter for the spot have the same shape in variation. The amplitude for SN mean RV is however slightly smaller.

Fig. 6 shows the correlations between the asymmetry parameters and the different estimates for the CCF barycenter (i.e. the SN mean RV, SN median RV and RV). The correlation between $\gamma$ and SN median RV is the strongest with a $R = 0.94$, followed by the correlation BIS SPAN - RV and $\gamma$ - SN mean RV, with $R = 0.86$. [[**Jessi:** Do the previous numbers need updating?]] Regarding the correlation between the width and the CCF barycenter, we note that the variation is seen as a circle in this parameter space and therefore no correlation is observed. Once again, like in the case of the facula, we see that some parameters of the SN gives stronger correlations than when using the Normal parameters.

As before, we corrected the originally RV's by using Eq. 5. The results of the correction are displayed in Fig. 7. Also in this case the proposed correction almost completely addresses the issue when considering the SN or Normal parameters, with val-

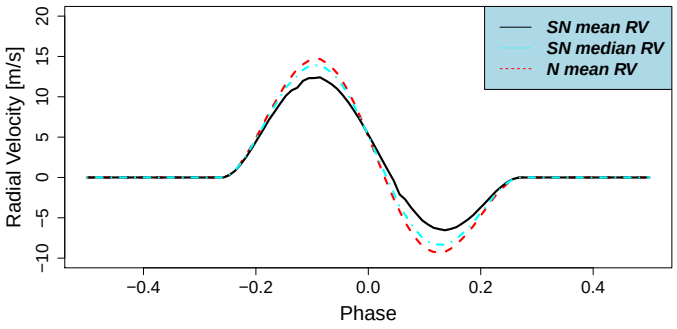


**Fig. 5.** RV's changes as function of the orbital phase in the case in which a spot is present on the photosphere of the star. SN mean RV seems to have the smallest spurious variations caused by the faculae.

| Parameter | N mean RV | SN mean RV | SN median RV |
|---|---|---|---|
| $\beta_0$ | 0.4975 | 0.21 | 0.21 |
| $\beta_1$ | $2e-16$ | $2e-16$ | $2e-16$ |
| $\beta_2$ | $2e-16$ | $2e-16$ | $2e-16$ |
| $\beta_3$ | 0.017 | 0.13 | 0.11 |
| $\beta_4$ | $2e-16$ | $2e-16$ | $2e-16$ |
| $R^2$ | 0.9959 | 0.9936 | 0.9952 |

**Table 3.** The p-values and coefficient of determination for the models fit using Eq. 5 to correct the estimated RVs from spurious variations caused by a spot. All the covariates are statistically significant to explain the variability in RV's caused by a spot, except the intercept and the width of the CCF. The estimated $R^2$ show that the proposed correction for stellar activity explains the vast majority of the spurious variability in the estimated RVs. [[**Jessi:** Isn't the width of the CCF parameter sig for N mean RV at 0.017? Previously it was noted that a 5% significance level was going to be used.]]

ues of $R^2$ larger than 0.99. [[**Xavier:** Looking at Fig. 7, we see that the activity correction proposed here is able to reduce the signal of a spot from a raw RV rms larger than 4.80 m s$^{-1}$down to a rms of 0.38 m s$^{-1}$. In this case the RV rms of the residuals obtained in the case of the Normal parameters is smaller, however,]] [[**Xavier:** REDO WITH FINAL PLOT with a difference of 6 $mm.s^{-1}$, we cannot say that this difference is significant. When comparing the activity correction proposed in this paper with what is commonly used, i.e only a linear dependance with the width and asymmetry of the CCF, we see that our solution is capable of reducing the RV residual rms by a factor of 3.5, which is even more than the factor 2 found in the case of the faculae.

In terms of the significance of the different parameters in Eq. 5, summmurized in Table 3, [[**Umberto:** we observe that the $\beta_0$, and $\beta_3$, corresponding respectively to the intercept and to the width of the CCF]], are [[**Umberto:** not helpful]] to understand the variation measured in RV's. This is not surprising when looking at the circle shape drawn when plotting the width as a function of the RV in Fig. 6. We see also that the ampltitude parameter, with coefficient $\beta_1$ is only useful to explain the RV variation of the spot in the case of the Normal distribution.]]

### 4.3. Spot and planet

The last simulation presented consists in having a planetary signal influencing the CCF, in addition to the 1% spot modeled before (see Sec. 4.2). [[**Xavier:** The purpose of this example is to check if we are able to disentangle as efficiently the two different sources of variations when using the parameters derived using a Normal, or a SN fit the the CCF. In this case, we inject a planet with an semi-amplitude of 10 m s$^{-1}$, with no eccentricity, and with a period corresponding to 1/3rd of the stellar rotational phase.]]

Fig. 8 shows the variation observed in the CCF barycenter parameters. Like in the case of the spot, all barycenter indicators show very similar variations, with SN mean RV showing a slightly smaller amplitude.

In Fig. 9, we show the correlation between the different CCF parameters. [[**Xavier:** Except of seing smaller correlation than in the case of the spot, due to the fact that the planet induces changes in the CCF barycenter without any change in any of the width or asymmetry parameters, the correlation strengths between the asymmetry parameters and the CCF barycenter are in exactly the same order than in the case of the spot: $\gamma$–SN median RV $R = -0.84$, BIS SPAN-RV $R = -0.78$ and $\gamma$–SN mean RV $R = -0.76$. The variation seen in the width-CCF barycenter space draw a circle like in the case of a spot, therefore, no correlation is observed between those parameters.]]

In order to correct the RV's from the spurious variation caused by the spot, [[**Xavier:** we need to add to our model of activity correction, a signal to take into account the RV variation caused by the injected planet. The observed RV can therefore be modeled by a combination of the activity and the planet signal:]]

$$RV = RV_{\text{activity}} + RV_{\text{planet}}, \tag{6}$$

[[**Xavier:** where $RV_{\text{activity}}$ can be found in Eq. 5, and $RV_{\text{planet}}$, in the case of no eccentricity, can be modelled by the following sinusoidal function:]]

$$RV_{\text{exoplanet}} = K \sin\left(\frac{2\pi}{P}(t - t_0)\right), \tag{7}$$

where the amplitude $K$, the orbital period $P$ and the epoch at the periapsis $t_0$ are three unknown parameters that define the planetary orbit.

The statistical tests conducted on the parameters, whose results are summarized in Table 4, [[**Xavier:** shows that except for the width parameters with coefficient $\beta_3$, all the other are significantly useful to explain the RV variation induce by a spot plus a planet.]]

### 4.4. Conclusions on the simulation study

In this Sec.4, we presented a first implementation of the SN fit to the CCF, using SOAP 2.0 to simulate noiseless CCF affected by stellar activity variation.

[[**Umberto:** Do we need to point out the following:
If we look at the $R^2$, the proposed function that corrects from stellar activity addresses (almost) all the spurious variations caused in RV's by active regions. Anyway, if we look the residuals, a systematic component is still present, although smaller respect the residual obtained with the usual correction. The residuals are supposed to be homoscedastic with 0 mean under the assumption that the model is correct. Therefore, since our $R^2$ is always close to 1, we can argue, as we discussed already many times, that the CCF does not follow a normal fit, neither a SN one, otherwise, given our extremely high $R^2$ we should have gotten homoscedastic 0 mean residuals.]]

[[**Xavier:** is all this discussion really usefull ? Before moving to real cases, where the analyses on five stars are presented,

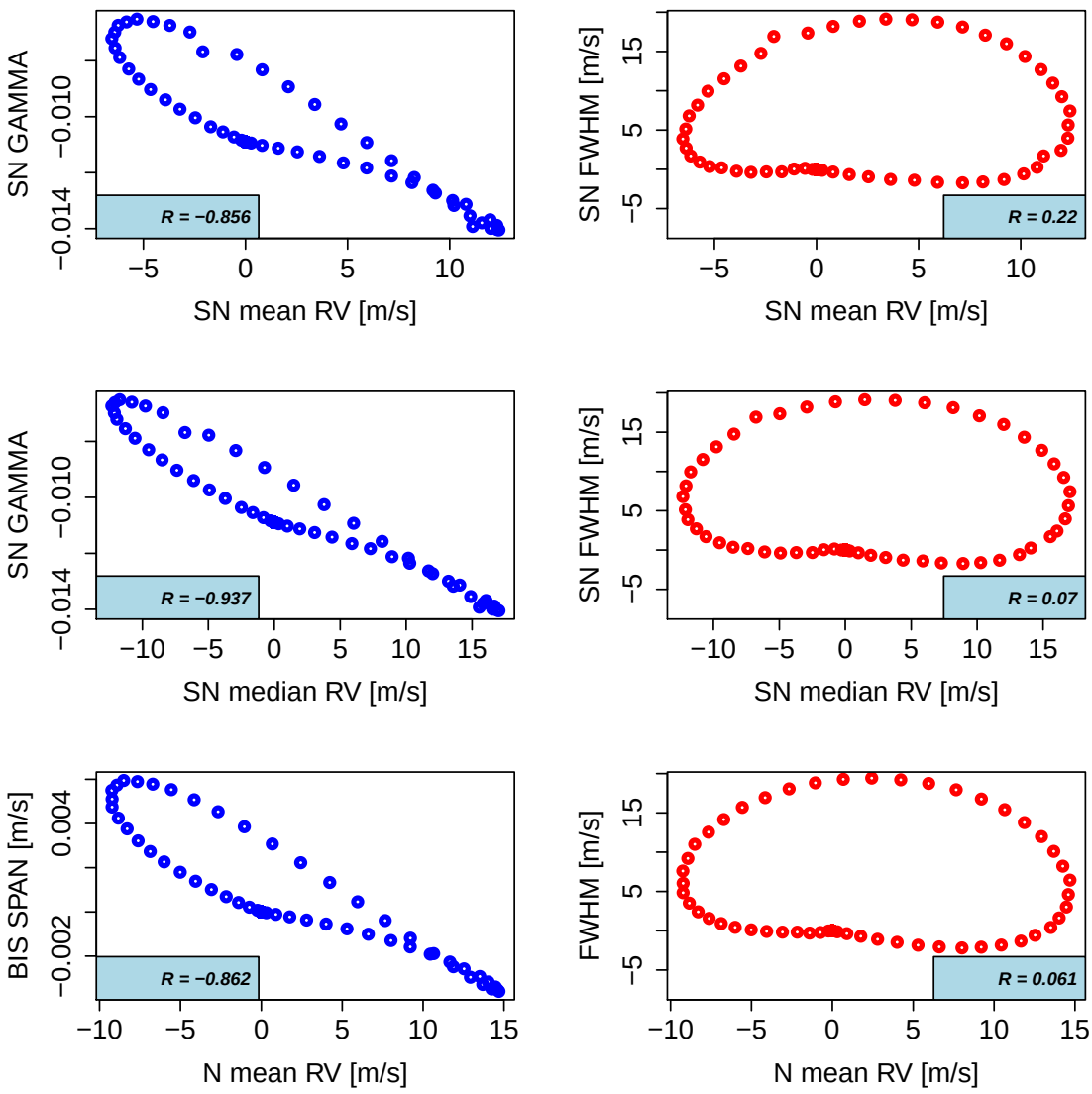


**Fig. 6.** Evaluation of the correlation between the RV's and the asymmetry parameters when a spot is present on the photosphere of the star. In this case only the shape of the CCF changes as the spot moves, producing statistically significative correlations only between the RV's and the asymmetry parameter.

| Parameter | N mean RV | SN mean RV | SN median RV |
|---|---|---|---|
| $\beta_0$ | 0.00063 | $2e-16$ | $1.42e-09$ |
| $\beta_1$ | $2e-16$ | $2e-16$ | $2e-16$ |
| $\beta_2$ | $2e-16$ | $2e-16$ | $2e-16$ |
| $\beta_3$ | 0.067 | 0.40 | 0.38 |
| $\beta_4$ | $2e-16$ | $2e-16$ | $2e-16$ |
| K | $2e-16$ | $2e-16$ | $2e-16$ |
| P | $2e-16$ | $2e-16$ | $2e-16$ |
| $t_0$ | $2e-16$ | $2e-16$ | $2e-16$ |
| Residuals | 0.71 m s$^{-1}$ | 0.66 m s$^{-1}$ | 0.70 m s$^{-1}$ |

**Table 4.** The p-values and coefficient of determination for the models fit using Eq. 6 to correct the estimated RVs from spurious variations caused by a spot while also accounting for a true planet RV. All the covariates are statistically significant to explain the variability in RV's caused by a spot, except the FWHM. Concerning the Keplerian parameters, the amplitude $K$ that provides relevance about the possibly presence of the exoplanet. Note that since nonlinear least squares was required, the residual standard error rather than the $R^2$ is displayed as a reference.

we need to provide further considerations. First of all, looking ad the analyses conducted with SOAP 2.0, it seems that the largest correlation between an asymmetry parameter and a set of RV's happens to be when respectively $\gamma$ and SN median RV are used. This is a bit surprising, since as the shape of the CCF changes, we expect SN median RV to be more robust than SN mean RV. [[**Umberto:** A possible justification of this ...]]. As second, when searching for stellar activity by deriving the correlation between the set of RV's and either an asymmetry parameter or the width of the CCF, the latter leads to weaker and hence less conclusive results if the active region is a spot. When stellar activity is dominated by faculae, both the shape and the width of the CCF changes as the faculae evolves on the photosphere of the star. Related to these last two considerations, we note that the interaction between the asymmetry and the width of the CCF is useful to explain part of the variability in the RV's if the active region is a spot but not when it is a faculae. The proposed function to correct for stellar activity addressed high level of spurious vari-

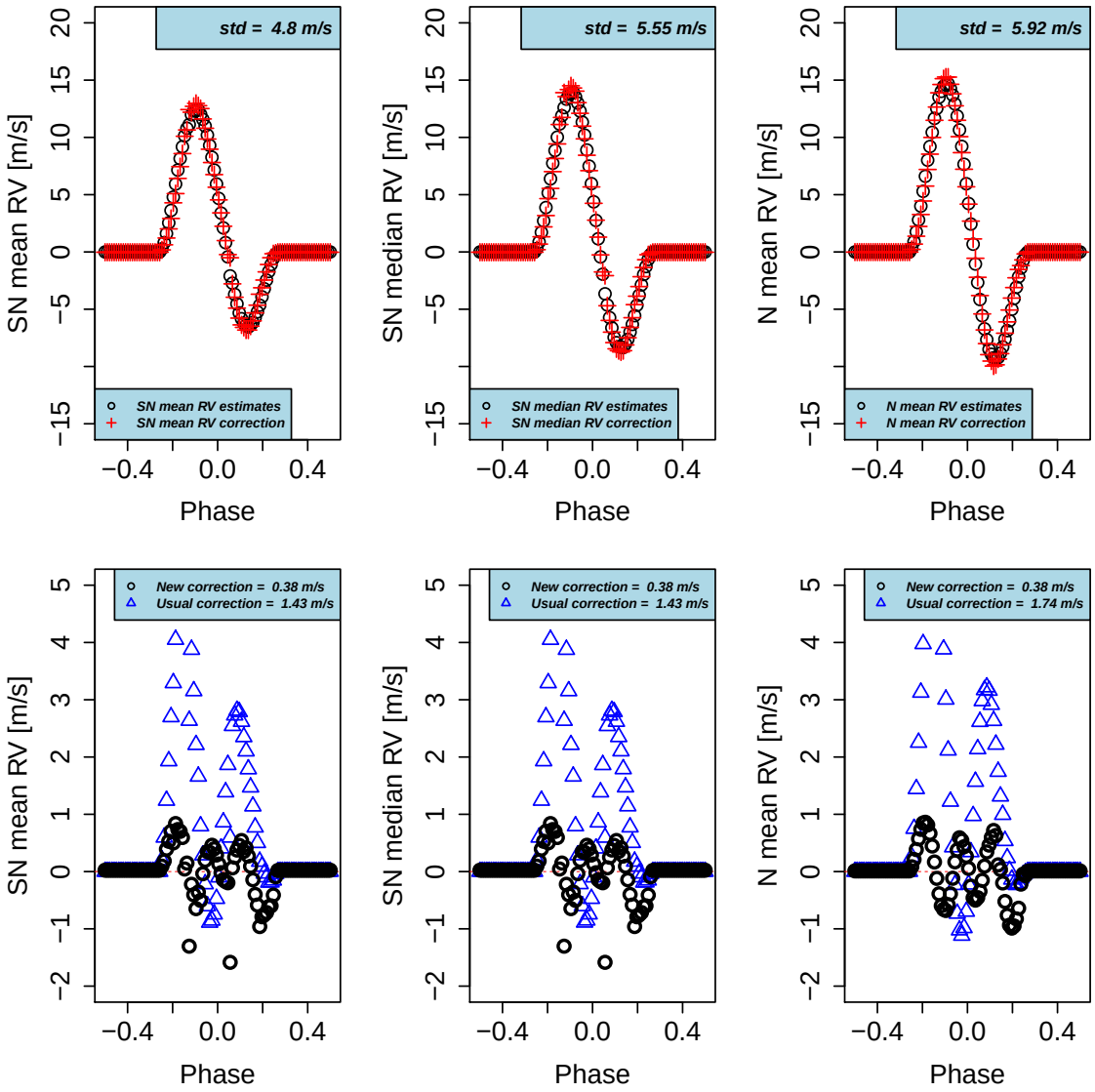


**Fig. 7.** [[**Umberto:** Set of spurious variations in RV's caused by a **spot** using a Normal and a SN fit before and once corrected from stellar activity. The new correction has done using Eq. 5 and the estimated parameters are presented in Table 2. Once corrected for stellar activity using the proposed linear function (black dots), the residuals present a smaller systematic component if compared with the residuals obtained with the usual correction (blue triangles).]]

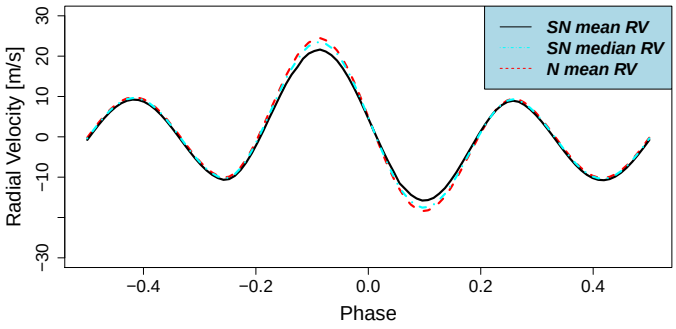


**Fig. 8.** RV's changes as function of the orbital phase in the case in which a spot is present on the photosphere of the star and a planet is injected. N mean RV seems to have the largest variations caused by the combined action of spot and planet.

ations in RV's caused by active regions. In particular, respect to other common linear interpolation, we proposed to use as covariates also the amplitude parameter of the CCF and the interaction between $\gamma$ and SN FWHM (or BIS SPAN and FWHM). As a consequence of using the interaction between the asymmetry and the width of the CCF, we note that the FWHM (or SN FWHM) becomes statistically not significant, while this is not the case if the interaction term is not involved in the linear regression. Finally, the correlations involving the common indicators (i.e. RV, FWHM and BIS SPAN) are systematically weaker than the correlations obtained by fitting the SN to the CCF, suggesting that this density could be helpful when searching for active regions. We recall moreover that all the quantities needed for conducting the analyses of the CCF are directly available by just fitting the SN.]]

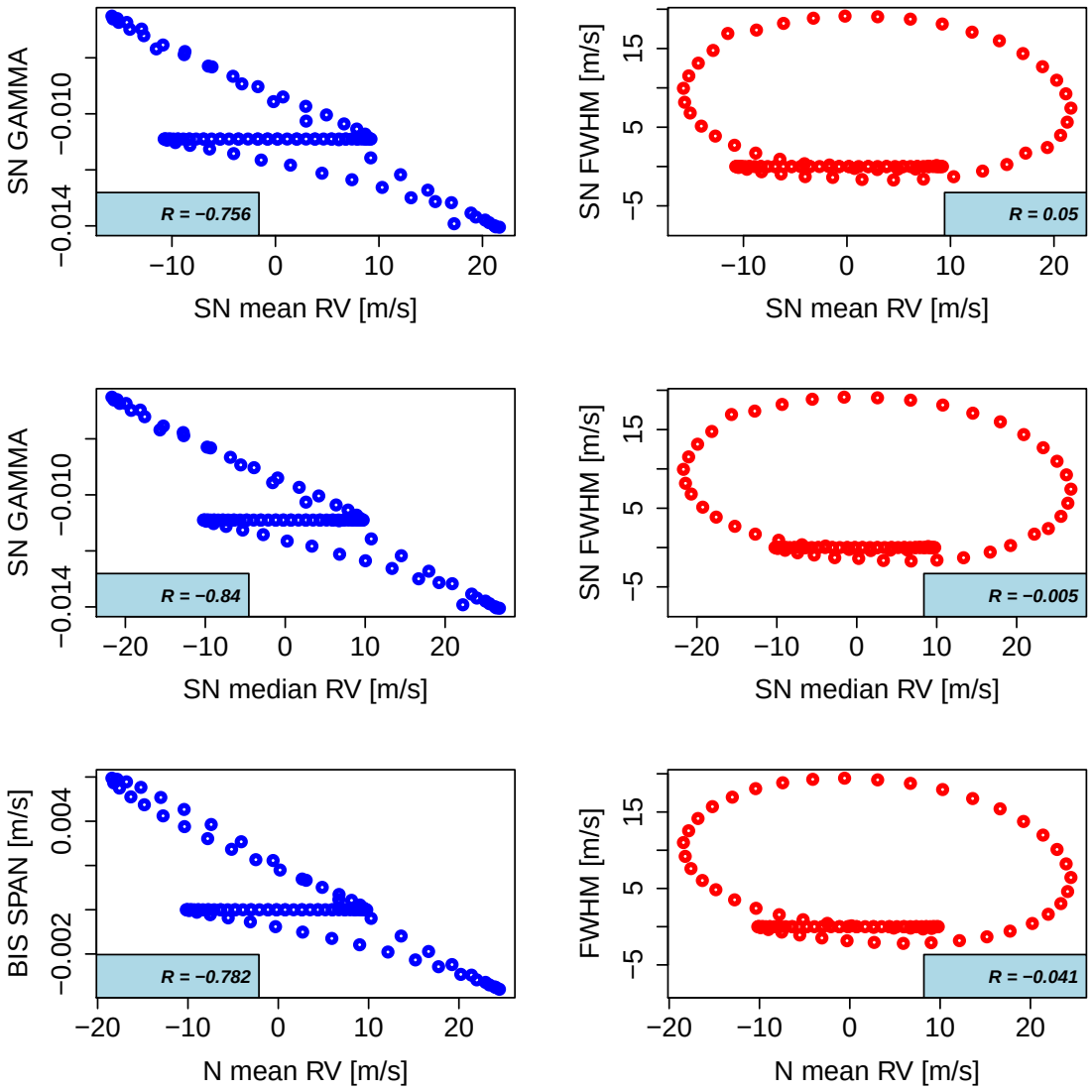


**Fig. 9.** Evaluation of the correlation between the RV's and the asymmetry parameters when a spot is present on the photosphere of the star and a planet is injected. In this case only the shape of the CCF changes as the spot moves, producing statistically significative correlations only between the RVs and the asymmetry parameter. The correlations between the RVs and the width parameter of the CCF is weaker than the previous case that considers only the presence of a spot on the photosphere of the star.

## 5. Real data application

In this Section we present the analyses conducted on Alpha Centauri b, comparing the result of fitting a CCF using the SN density defined in Sec. 2.1 with the approach based on fitting a Normal density to retrieve the RV and width of the CCF and calculating the bisector to derive the asymmetry parameter BIS SPAN. Four other stars have been analyzed with the proposed method and details can be found in the Appendix A. For all the stars that have been considered in the present work, [[**Xavier:** we selected CCFs that were derived from spectra that had a SNR at 550 nm larger than 10.]]

### 5.1. Comparison between the different CCF parameters derived with the Normal and the Skew Normal density

We analysed the 1808 CCFs derived from spectra of Alpha Centauri B taken in 2010 by the HARPS spectrograph. Note that more observations have been done during this year, however we selected here only the data that are not significantly affected by contamination from Alpha Cen A (see Dumusque et al. 2012). [[**Xavier:** We choose these observations as their represent probably the best sampled and most precise RV data set showing a strong solar-like activity signal (Thompson et al. 2017; Dumusque et al. 2012).]]

We begin the analyses by evaluating the correlation between $\gamma$ and the BIS SPAN. In the left panel of Fig. 10, we see that the relationship between $\gamma$ and the BIS SPAN is linear, with a slope equal to 0.72 and a strong Pearson correlation coefficient of $R = 0.95$. This strong correlation shows that $\gamma$ and BIS SPAN are measuring in a very similar way the CCF asymmetry. [[**Xavier:** This strong correlation allows as well to convert the dimensionless $\gamma$ parameters in $\mathrm{m\,s^{-1}}$to compare it with BIS SPAN using the slope of the correlation, in this case 720 $\mathrm{m\,s^{-1}}$.]]

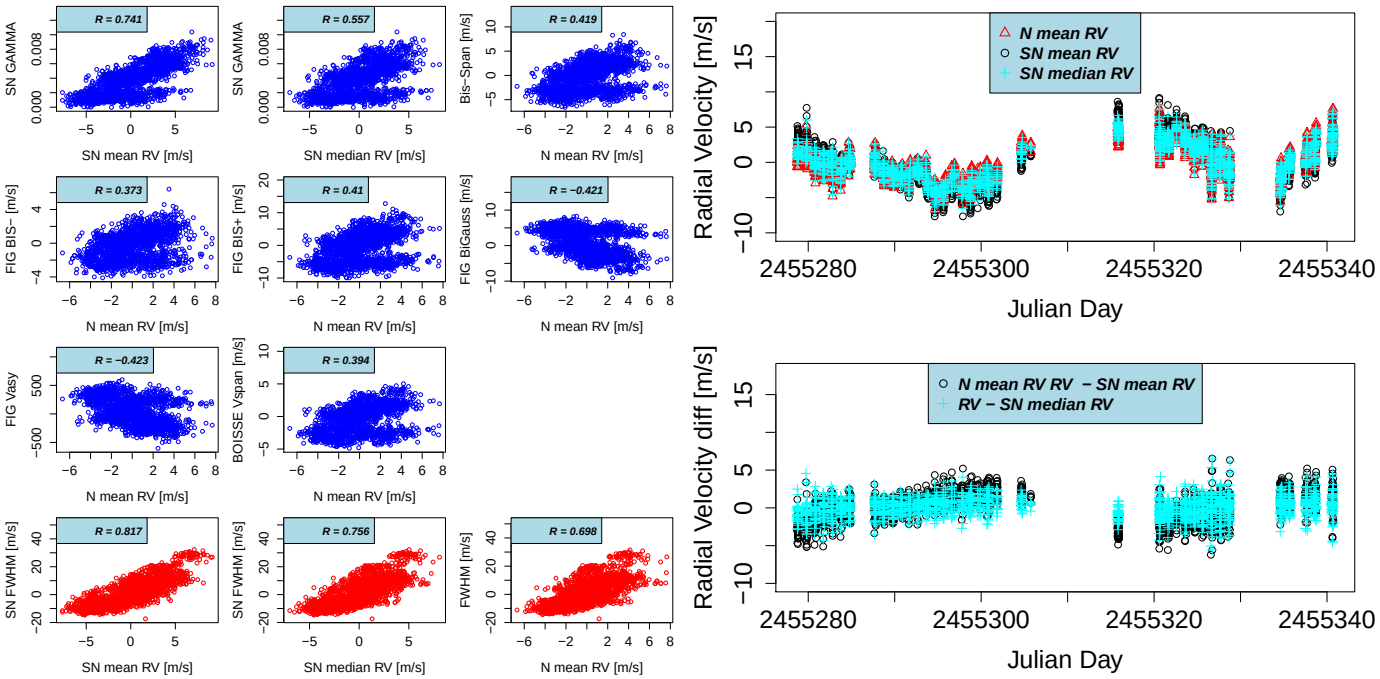


**Fig. 10.** *Left:* Correlation between $\gamma$ and the BIS SPAN for Alpha Centauri B. As we can see with the strong correlation, those two parameters measure the CCF asymmetry in the same way. *Top right:* RVs as function of Julian Day for Alpha Centauri b in 2010. The RVs are retrieved using the mean of a Normal fitted to the CCF (red triangles), or the mean (black circles) or median (cyan crosses) of a SN density fitted to the CCF. *Bottom right:* RV differences between the RV derived when using the SN density and the RV derived when using the Normal density.

In the right of Fig. 10, we show the comparison between the RVs retrieved using the SN density and the ones obtained with the Normal density. [[**Xavier:** We clearly see that the amplitude of the activity signal is stronger when using SN mean RV, and the signal measured using N mean RV or SN median RV are very similar. We observe the same behavior in the case of a faculae (see Sec. 4.1) and therefore it seems that the activity signal seen in those data are more likely affected by faculae. This is an observation that was already made in Dumusque (2014).]]

[[**Xavier:** Similar to what is done in Sec. 4, we compare the correlation between the asymmetry or the width parameters of the CCF and the RV, i.e. the CCF barycenter, in Fig. 11. In addition to what is done in Sec. 4, we add the asymmetry parameters derived in Boisse et al. (2011), $V_{span}$ and in Figueira et al. (2013), BIS-, BIS+, Bi Gauss and $V_{asy}$, as these authors found those asymmetry parameters more correlated to the RVs than BIS SPAN. It is clear in the case of Alpha Cen B, but also in the four other stars presented in Appendix A, that the correlation found between $\gamma$ and SN mean RV is always the strongest. On alpha Cen B, the Pearson correlation coefficient reaches a value of $R = 0.74$, while it reaches at best $R = -0.42$ for all the other asymmetry-RV correlations not derived using the SN density fit. When looking at the width-RV correlation, once again, the correlation between SN FWHM and SN mean RV is the strongest for all the stars, except for the quietest star of all five, HD10700, for which the Pearson correlation coefficient 0.53 for the Normal parameters and 0.42 for the SN parameters. This analysis shows that the parameters derived when using a SN density are more sensitive the stellar activity, and therefore using those parameters can lead to the detection of stellar activity, when the Normal parameters do not detect anything (see the Appendix A the case of the the asymmetry-RV correlation for HD10700, HD215152, Corot-7, and the width-RV correlation for HD215152)]]

[[**Xavier:** In Fig 12, we study the efficiency of the stellar activity correction proposed in Sec. 3. The first observation is that the RV measured with SN mean RV presents a rms 35% larger than the RV measured when fitting a Normal density. In the case of SN median RV, we see as well a larger RV rms, however this one is only 9% larger. Even though we see those differences in RV, once we correct for the stellar activity using Eq. 5, we find the same RV residuals rms. In the best case, for SN mean RV, we reduce the stellar activity signal by a factor of 2, while in the worst case, N mean RV, only a factor 1.5. Therefore, although it seems that the parameters derived using the SN density are more sensitive to stellar activity, we are not able to correct better for stellar activity signal using a linear combination of the different CCF parameters.]]

[[**Xavier:** When looking at the significance of the different parameters in the activity correction in Table 5, we see that the intercept is not significant for any of the correction, and BIS SPAN, with coefficient $\beta_2$, is not useful to explain stellar activity when using the parameters derived from the Normal density fit. However all the other parameters in the Normal and SN case are useful. By analyzing the value for $R^2$, we see that the correction is more efficient at explaining stellar activity when considering for the CCF barycenter SN mean RV. This is expected from the results above since for the three different cases, we arrive to the same RV residual rms after correction, but before correction SN mean RV shows the largest RV rms.]]

### 5.2. Detection limits when using the RV derived with the Normal or the Skew Normal density

[[**Xavier:** In the preceding section, we saw that the RV measured when considering a SN or a Normal density present different amplitudes, mainly in the case of SN mean RV. However,

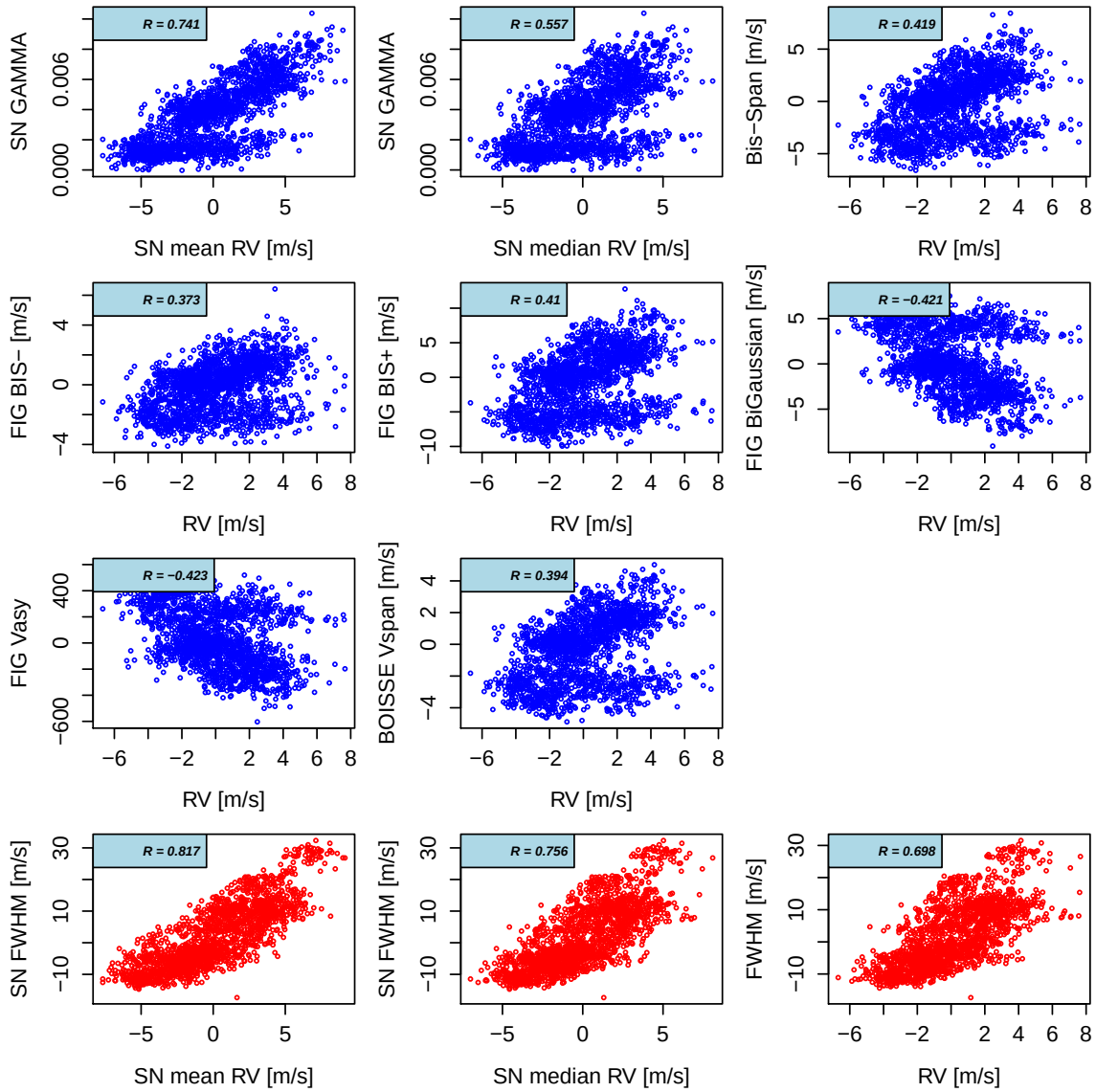


**Fig. 11.** Correlation between the asymmetry parameters and the RVs for Alpha Centauri B. The last three plots show the correlation between the FWHM and the RVs for Alpha Centauri B using respectively the SN (SN mean RV and SN median RV) and the Normal fits. The correlations are always stronger when using parameters derived from the SN fit than the Normal one. The p–values associated with each $R$ is statistically different from 0.

when correcting for stellar activity using the linear combination presented in Eq. 5, it seems that we are able to reach the same RV precision. In this section, we want to investigate if for some of the three different RV definitions presented in this paper, it is easier or more difficult to detect planetary signals when considering stellar activity. To do so, we estimated what would be the minimum amplitude of a planetray signal we could detect at different orbital periods.

To get CCF affected by realistic stellar activity signals, we took the ones from Alpha Cen B used in the preceding section. To simulate a planetary signal, when then blue- or red-shifted those CCF with the wanted amplitude, period and phase.

We first simulated several RV data sets, with always the same stellar signal, but injecting planets with parameters corresponding to the following grid:

- period of 3, 5, 7, 9, 11, 15, 20 , 25 and 30 days,
- amplitude from 0.5 to 3 $\mathrm{m\,s^{-1}}$ by steps of 0.05 $\mathrm{m\,s^{-1}}$,
- 10 different phases, evenly sampled between 0 and $2\pi$.

For each simulation, 4500 in total, we applied the activity correction presented in Eq. 5 and look for signals in the residuals using a Generalized Lomb-Scargle periodogram (GLS Zechmeister & Kürster 2009). If a signal with a p–value[5] smaller than 1% had a period compatible with the injected planetary period considering an error of 20%, we were estimating this signal as significant, and therefore the corresponding planet as detected.

Finally, for each period, we searched for the minimum amplitude for which at least 50% of the planets with different phases were detected. The results are shown in Fig. 13. As we can see, all the different RV estimates gives extremely similar detection limits. Therefore, we can use any of these estimates when searching for planetary signal in RV data contaminated by stellar activity.

[5] estimated by bootstrap

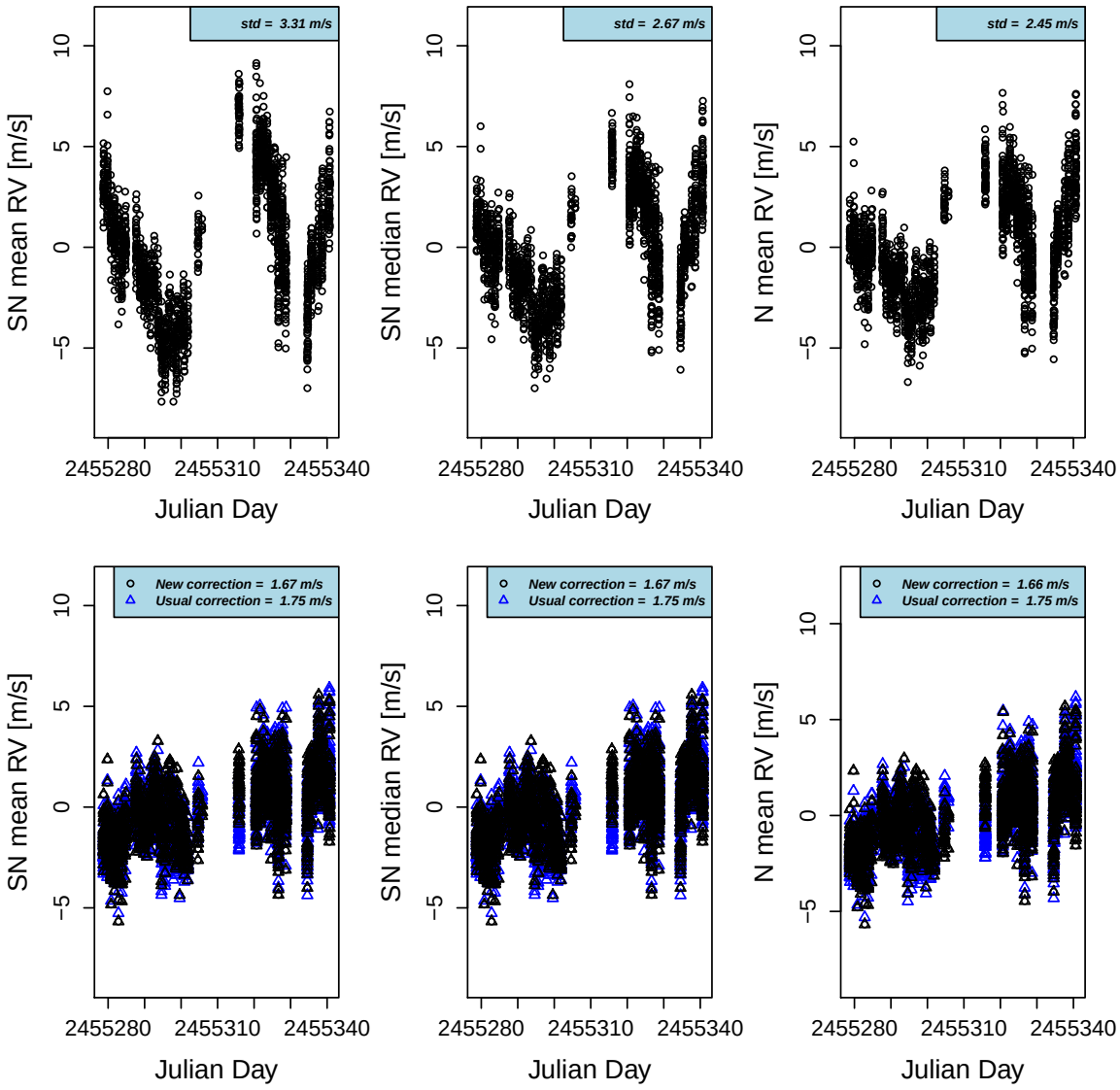


**Fig. 12.** Set of RV's for Alpha Centauri B using a Normal and a SN fit before and once corrected from stellar activity. The correction is done using Eq. 5. Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable.

| Parameter | RV | SN mean RV | SN median RV |
|---|---|---|---|
| $\beta_0$ | 0.49 | 0.90 | 0.027 |
| $\beta_1$ | $2.22e-16$ | $2.22e-16$ | $2.22e-16$ |
| $\beta_2$ | 0.33 | $2.22e-16$ | $1.23e-11$ |
| $\beta_3$ | $2.22e-16$ | $2.22e-16$ | $2.22e-16$ |
| $\beta_4$ | $2.22e-16$ | $2.22e-16$ | $2.22e-16$ |
| $R^2$ | 0.57 | 0.78 | 0.66 |

**Table 5.** Evaluation of the linear combination used for correcting the RV's, according to Eq. 5. Concerning the Normal fit, all the parameters but the intercept and the BIS SPAN are useful in explaining variations in RV's of the star that can be caused by stellar activity. For the SN fit we note that the parameter related to $\gamma$ is highly significant to address part of the spurious variations in RV's caused by stellar activity. The evaluation of the $R^2$ shows that the proposed linear combination better explains variations in RV's due to stellar activity coming from the SN analysis that uses SN mean RV.

Although the detection limits stay constant for all the periods below 15 days, we see that for 20 and 30 days, we have a huge increase in detection limits for all the RV estimates. This is because period of the simulated planet, 20 days and 30 days, are close to the first harmonic or to the rotational period of the star (36.7 days, Dumusque et al. 2012) and therefore close to the semi-periodicity of the stellar activity signal. In such a case, the activity correction absorbs part of the planetary signal. We thus need to inject a much stronger planetary signal at these periods to detect them.]]

## 6. Estimation of standard errors for the CCF parameters

[[**Xavier:** In this section, we investigate how the noise of the CCF influence the CCF parameters derived either by a Normal density or a SN density fit. Because a CCF is obtained from a cross-correlation, each point in a CCF is correlated with each other. Therefore, we cannot simply vary each point in the CCF with their respective error bar and then recalculate the best SN or Normal density fit, to see how the CCF noise influences N

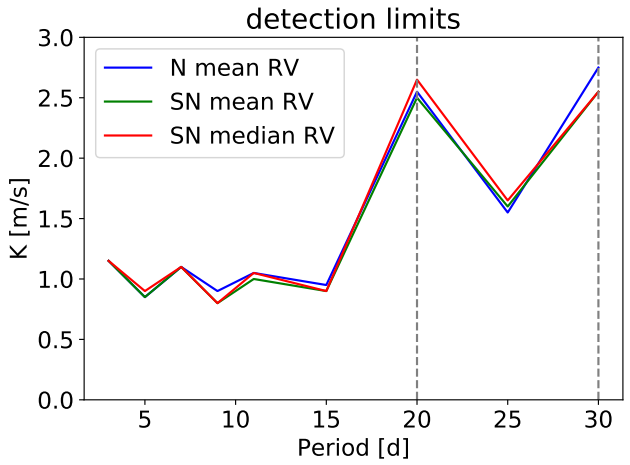


**Fig. 13.** Detection limits of planetary signals once we have removed from the raw RVs the linear combination of CCF parameters proposed in Eq. 5 to correct for stellar activity. As we can see, when considering the different parametrization for the RV: N mean RV, SN mean RV and SN median RV, we arrive to very similar detection limits.

mean RV, SN mean RV, SN median RV, FWHM, SN FWHM, BIS SPAN and $\gamma$. We therefore started from the spectrum, for which we know that each point is totally independent of the others. The standard error on each point of the spectrum is given by the photon noise, which follows a Poisson law and is therefore obtained by taking the square root of the measured flux.]]

[[**Xavier:** We applied the following method to estimate the error bars on the different parameters derived from the CCF. We first modify the values of all the points in the spectrum given their respective error bars. To do so we added to each point the value randomly drawn from a Gaussian distribution centered on the value of the point and with standard deviation the square root of the flux. We then calculated the CCF from this spectrum using the method presented in Pepe et al. (2002), fitted either a Normal or SN density to this CCF and recorded the different parameters. We redo this process a hundred times, which gives us at the end, a distribution for each CFF parameter. The standard deviations of the obtained distributions is then associated to the noise of each CCF parameters.]]

[[**Xavier:** We measured the standard deviation of each CCF parameters for all the CCF of HD215152, HD192310 and Corot-7. This gives us an information of how the noise on each CCF parameter varies as a function of SNR. In this case, the SNR, measured at 550 nm on the original spectra, varies between 10 and 500. The results can be seen in Fig. 14. ]]

### 6.1. Estimation of standard errors for the CCF parameters for real stars

In the top plots of Fig. 14 we show the different errors for the RVs, either defined as RV (red triangles), SN mean RV (black circles) or SN median RV (cyan crosses), the width and the asymmetry of the CCFs for three star, HD215152, HD192310 and Corot-7, that are all at different SNR levels. The parameter SN50 corresponds to the SNR in order 50, which defines a wavelength of 550 nm. In the bottom plots, we show the ratio between the parameters derived from the bootstrap analysis fitting the SN and the parameters derived from the bootstrap analysis fitting the Normal density. We first see that the errors on the CCF parameters only depends on the SNR and do not depend on the spectral type. This is true if the spectral type are not too different though, like here where we show the results for G and K dwarfs.

Concerning the standard errors related to the RVs, the ratio between the RV error measured by the bootstrap using the SN and Normal fitting is 1.6 when using SN mean RV and 0.9 when using SN median RV. In other words, by using SN median RV as parameter that defines the radial velocity of the star given a CCF, we get standard errors 10% smaller than using the Normal fit and its corresponding mean. This result is consistent with what we observed with the simulation from SOAP presented in Sec. 4.

Regarding the errors in width of the CCF, we see that the bootstrap analysis for the Normal and the SN are comparable. Therefore, the precision in the width of the CCF is the comparable if we fit a Normal or a SN to the CCF.

Finally, for the errors in evaluating the asymmetry of the CCF, we see that, when fitting the SN to the CCF, the asymmetry errors are 15% smaller. Therefore, the SN fit gives a better precision in CCF asymmetry than what can be reached using BIS SPAN. We recall moreover that, using the SN, all parameters are automatically retrieved in 1 single step, while in the common approach the RV and the BIS SPAN are calculated separately.

## 7. Discussion

An analysis of the CCF residuals after fitting a Normal or SN density shows that the SN is a slightly better model to explain the shape of the CCF. This comes from the fact that CCF's present a natural asymmetry due the convective blueshift.

We tested at first our assumptions by using simulated CCF's retrieved using the software SOAP 2.0. We then compared for five real stars the difference between the RV's (defined as mean of a Normal, mean of the SN or median of the SN), FWHM and asymmetry (BIS SPAN in the Normal case and $\gamma$ in the SN case). The $\gamma$ parameter is linearly dependent on the BIS SPAN, with always a strong correlation coefficient. The slope of this linear correlation changes depending on the studied star. This is probably because the spectral type is different, therefore the effects from stellar activity are different.

When using as parameter for the RV the mean of the SN, the standard errors are in average 60% larger than the standard errors retrieved fitting a Normal. However, once the RV is defined as the median of the SN (cyan crosses), the standard errors are in average 10% smaller than the standard errors coming from the Normal fit. When looking at the correlation between the asymmetry and width parameters of the CCF (FWHM and BIS SPAN or the alternative indicators in Figueira et al. (2013) in the Normal case, and SN FWHM and $\gamma$ in the SN case) with respect to the RV's (RV's in the Normal case or SN RV's in the SN case), we observe that the correlations are always stronger for the parameters of the SN. Therefore, the SN parameters are more sensitive to activity. In the case of Tau Ceti, which is at very low activity level, we find a significant correlation of 0.322 between $\gamma$ and SN mean RV, while for all the other asymmetric parameterization, BIS SPAN or the alternative indicators in Figueira et al. (2013), the correlations are weaker with a maximum of 0.225.

## 8. Conclusion

In this paper we introduced a novel approach based on the Skew Normal (SN) density to retrieve RV's and shape variations in the CCF of stars. When searching for small-mass exoplanets using

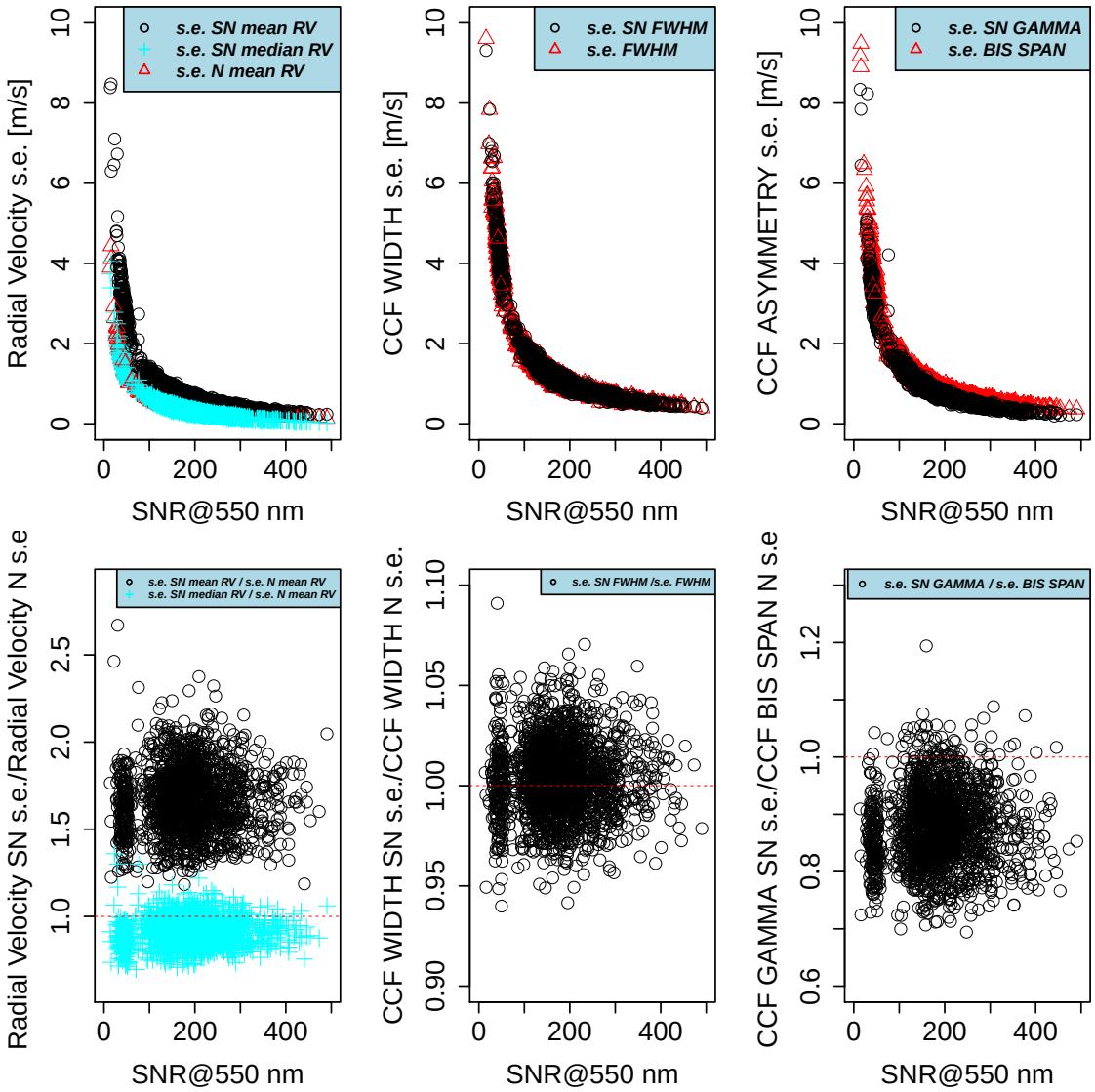


**Fig. 14.** Comparison between the standard errors using the bootstrap analysis for the RVs, the FWHM and the asymmetry parameter. When using SN mean RV (black circles), the standard errors are in average 60% larger than the standard errors retrieved fitting a Normal (red triangles). However, if using SN median RV (cyan crosses), the standard errors are in average 10% smaller than the standard errors coming from the Normal fit. To use as asymmetry parameter $\gamma$ of the SN leads to standard errors in average 15% smaller than the standard errors related to the BIS SPAN. Note that for the asymmetry, the error in BIS SPAN is in $\mathrm{km\,s^{-1}}$. To be able to compare the errors in $\gamma$ and BIS SPAN, we multiplied the error in $\gamma$ by the slope of the correlation between $\gamma$ and BIS SPAN.

the RV technique, it is essential to understand variation of the shape of the CCF, which is a proxy for stellar activity effects. The standard approach consist at first to fit a Normal density to the CCF to get the RV and FWHM, defined as the mean and the FWHM of the Normal density, and then to measure the asymmetry of the CCF by calculating the BIS SPAN. FWHM and BIS SPAN give us information on the line shape that are used to probe stellar activity signals.

In this paper we propose to conduct the analysis fitting a SN density to the CCF. Since the CCF presents a natural asymmetry due the convective blueshift, the SN density can in principle better catch these aspects respect the Normal fit. On top of that, by using the SN density to fit CCF's, we can measure simultaneously the RV of the star, the width and the asymmetry of the CCF.

Starting from the simulation environment SOAP and then moving to real stars, we showed that using the SN density to fit CCF's leads to a significant improvement to probing stellar activity. While for the Normal density mean and median are equivalent, using the SN fit different location parameters can be tested. While the median of the SN is a more robust statistic respect variations in the shape of the CCF, the mean of the SN is more sensible to changes in the asymmetry of the CCF. We suggest to use as parameter that defines the RV of the star the median of the SN, since the standard errors related to this parameter working with CCF's from real stars are on average 10% smaller than the standard errors retrieved using the Normal density. To evaluate changes in the asymmetry of the CCF, we suggest to use the mean of the SN. The correlation between SN mean RV and SN FWHM and the correlation between SN mean RV and $\gamma$ (the asymmetry parameter of the SN) are much stronger than

the correlations between the equivalent parameters derived using a Normal fit (RV, FWHM and BIS SPAN or the asymmetric parameters described in Figueira et al. (2013)). The precision on the asymmetry measured by $\gamma$ is greater than the one on BIS SPAN by ∼15%. Therefore when searching for rotational periods in the data, or applying Gaussian Processes to account for stellar activity signals, the SN parameters should be used.

Because of stellar activity the estimated RV's are contaminated by spurious variations. We propose to use a function that corrects from stellar activity that beyond the width and the asymmetry parameters of the CCF includes also the contrast parameter A and the fourth parameter defined as the interaction between width and the asymmetry parameters of the CCF. We found these new two parameter helpful to explain part of the spurious variations in RV's caused by stellar activity.

Finally, we also encourage the use of bootstrapping to estimate more realistic errors on the different parameters of the Normal or SN fitted to the CCF, mainly in the low SNR regime where a gain of 50% can be reached. This takes significantly more time, but note that 100 bootstrapped dataset are enough to get a good estimation of errors.

## 9. Acknowledgements

We are grateful to all technical and scientific collaborators of the HARPS Consortium, ESO Headquarters and ESO La Silla who have contributed with their extraordinary passion and valuable work to the success of the HARPS project. XD is grateful to The Branco Weiss Fellowship–Society in Science for its financial support.

## Appendix A: Appendix

In this Appendix we present the analyses conducted on other 4 stars: HD192310, HD10700, HD215152 and finally Corot-7.

[[**Umberto:** Add further information about the stars here presented.]][[**Xavier:** ]]

Table A.1 summarizes the results obtained by the SN fit and the some of the results based on the Normal fit. The results are all consistent with the conclusions derived by the analyses on Alpha Centauri b. The correlation between $\gamma$ and SN mean RV is stronger than the correlation between the BIS SPAN and RV for all the considered stars. The correlation between SN FWHM and SN mean RV is stronger than the correlation between FWHM and RV for three of the four stars. Also for all these stars we corrected the originally estimated RV's from spurious variations in RV's caused by stellar activity, using Eq. 5. Fig. A.2–A.6 show the resulting corrected RV's. While the Normal and SN residuals, once corrected for stellar activity, are comparable for the stars HD192310 and HD10700, the results of the analyses on the star HD215152 (whose CCF's have lower SNR respect to the previous two analyzed stars) suggest that the residuals for the Normal are 0.054 m s$^{-1}$ higher than the residuals retrieved with the SN analysis. Finally, the results of the analysis on Corot 7, whose CCF's have lowest SNR, show that once corrected from stellar activity the residuals from the Normal fit are 0.336 m s$^{-1}$ higher than the residuals retrieved with the SN analysis.

| Star | # CCFs | R(SN $\gamma$, Bis-Span) | slope(SN $\gamma$, Bis-Span) | R(SN $\gamma$, SN mean RV) | R(Bis-Span, RV) | R(FIG BiGaussian, RV) | R(S |
|---|---|---|---|---|---|---|---|
| HD192310 | 1577 | 0.888 | 0.786 | 0.669(0.64; 0.695) | 0.329(0.285; 0.373) | −0.333(−0.376; −0.289) | |
| HD10700 | 7928 | 0.78 | 0.604 | 0.323(0.302; 0.342) | −0.073(−0.095; −0.0051) | 0.127(0.105; 0.148) | |
| HD215152 | 273 | 0.763 | 0.794 | 0.571(0.485; 0.646) | −0.067(−0.184; 0.052) | 0.269(0.155; 0.376) | |
| Corot 7 | 173 | 0.814 | 0.607 | 0.561(0.450; 0.656) | 0.092(−0.058; 0.238) | −0.335(−0.228; −0.082) | |

**Table A.1.** Subset of notable correlations between the asymmetry parameter (and the FWHM) and the RVs for four stars: HD192310, HD10700, HD215152 and Corot 7. The complete results of the analyses of the correlations for the four stars are presented in Fig. A.1–A.7.

## References

Anglada-Escudé, G. & Butler, R. P. 2012, The Astrophysical Journal Supplement Series, 200, 15
Arellano, R. B. & Azzalini, A. 2010
Azzalini, A. 1985, Scandinavian journal of statistics, 171
Azzalini, A. & Capitanio, A. 2014, The skew-normal and related families. Institute of Mathematical Statistics Monographs
Baranne, A., Queloz, D., Mayor, M., et al. 1996, Astronomy and Astrophysics Supplement Series, 119, 373
Belsley, D. A. 1991, Conditioning diagnostics: Collinearity and weak data in regression No. 519.536 B452 (Wiley New York)
Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, Astronomy & Astrophysics, 528, A4
Boisse, I., Moutou, C., Vidal-Madjar, A., et al. 2009, Astronomy & Astrophysics, 495, 959
Claret, A. & Bloemen, S. 2011, A&A, 529, A75
Cosentino, R., Lovis, C., Pepe, F., et al. 2012, in Proc. SPIE, Vol. 8446, 84461V
Desort, M., Lagrange, A.-M., Galland, F., Udry, S., & Mayor, M. 2007, Astronomy & Astrophysics, 473, 983
Dumusque, X. 2014, ApJ, 796, 133
Dumusque, X. 2016, Astronomy & Astrophysics, 593, A5
Dumusque, X., Boisse, I., & Santos, N. 2014, The Astrophysical Journal, 796, 132
Dumusque, X., Borsa, F., Damasso, M., et al. 2017, Astronomy & Astrophysics, 598, A133
Dumusque, X., Pepe, F., Lovis, C., et al. 2012, Nature, 491, 207
Dumusque, X., Udry, S., Lovis, C., Santos, N. C., & Monteiro, M. 2011, Astronomy & Astrophysics, 525, A140
Efroymson, M. 1960, Mathematical methods for digital computers, 191
Feng, F., Tuomi, M., & Jones, H. R. 2017, Astronomy & Astrophysics, 605, A103
Figueira, P., Santos, N., Pepe, F., Lovis, C., & Nardetto, N. 2013, Astronomy & Astrophysics, 557, A93
Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, Publications of the Astronomical Society of the Pacific, 128, 066001
Hatzes, A. P. 2002, Astronomische Nachrichten, 323, 392
Haywood, R., Collier Cameron, A., Queloz, D., et al. 2014, Monthly notices of the royal astronomical society, 443, 2517
Hocking, R. R. 1976, Biometrics, 32, 1
Kurster, M., Endl, M., Rouesnel, F., et al. 2003, ASTRONOMY AND ASTROPHYSICS-BERLIN-, 403, 1077
Lagrange, A.-M., Desort, M., & Meunier, N. 2010, Astronomy & Astrophysics, 512, A38
Lindegren, L. & Dravins, D. 2003, Astronomy & Astrophysics, 401, 1185
Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
Meunier, N., Desort, M., & Lagrange, A.-M. 2010, Astronomy & Astrophysics, 512, A39
Oshagh, M., Boisse, I., Boué, G., et al. 2013, A&A, 549, A35
Pepe, F., Mayor, M., Galland, F., et al. 2002, Astronomy & Astrophysics, 388, 632
Pepe, F., Molaro, P., Cristiani, S., et al. 2014, Astronomische Nachrichten, 335, 8
Queloz, D., Bouchy, F., Moutou, C., et al. 2009, Astronomy & Astrophysics, 506, 303
Queloz, D., Henry, G., Sivan, J., et al. 2001, Astronomy & Astrophysics, 379, 279
Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, Monthly Notices of the Royal Astronomical Society, 452, 2269
Robertson, P., Mahadevan, S., Endl, M., & Roy, A. 2014, Science, 1253253
Saar, S. H. & Donahue, R. A. 1997, The Astrophysical Journal, 485, 319
Thompson, A., Watson, C., de Mooij, E., & Jess, D. 2017, Monthly Notices of the Royal Astronomical Society: Letters, 468, L16
Zechmeister, M. & Kürster, M. 2009, A&A, 496, 577

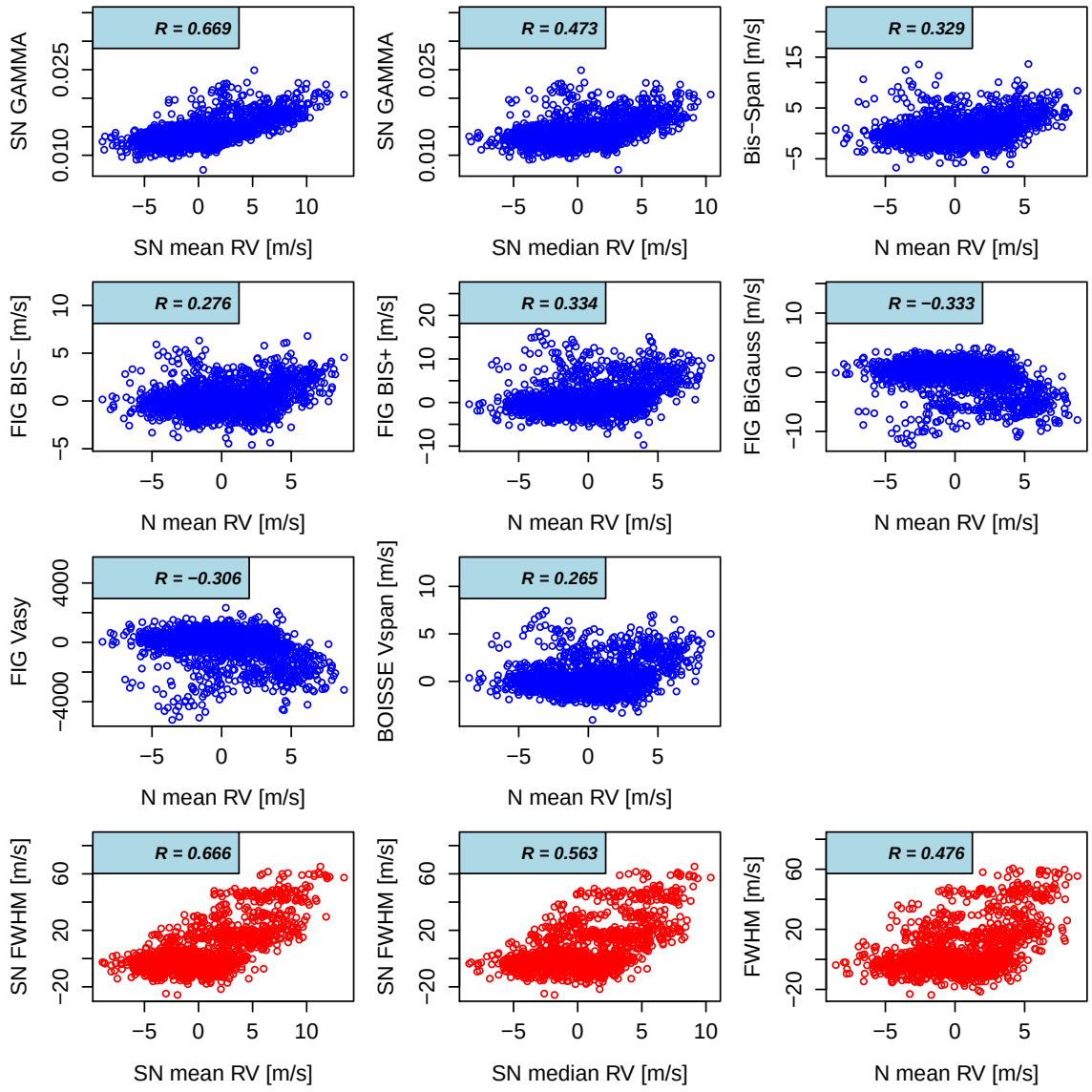


**Fig. A.1.** Correlation between the asymmetry parameters and the RV's for HD192310. The last three plots show the correlation between the FWHM and the RV's for HD192310 using respectively the SN and the Normal fits. The p–values associated with each $R$ is statistically different from 0.

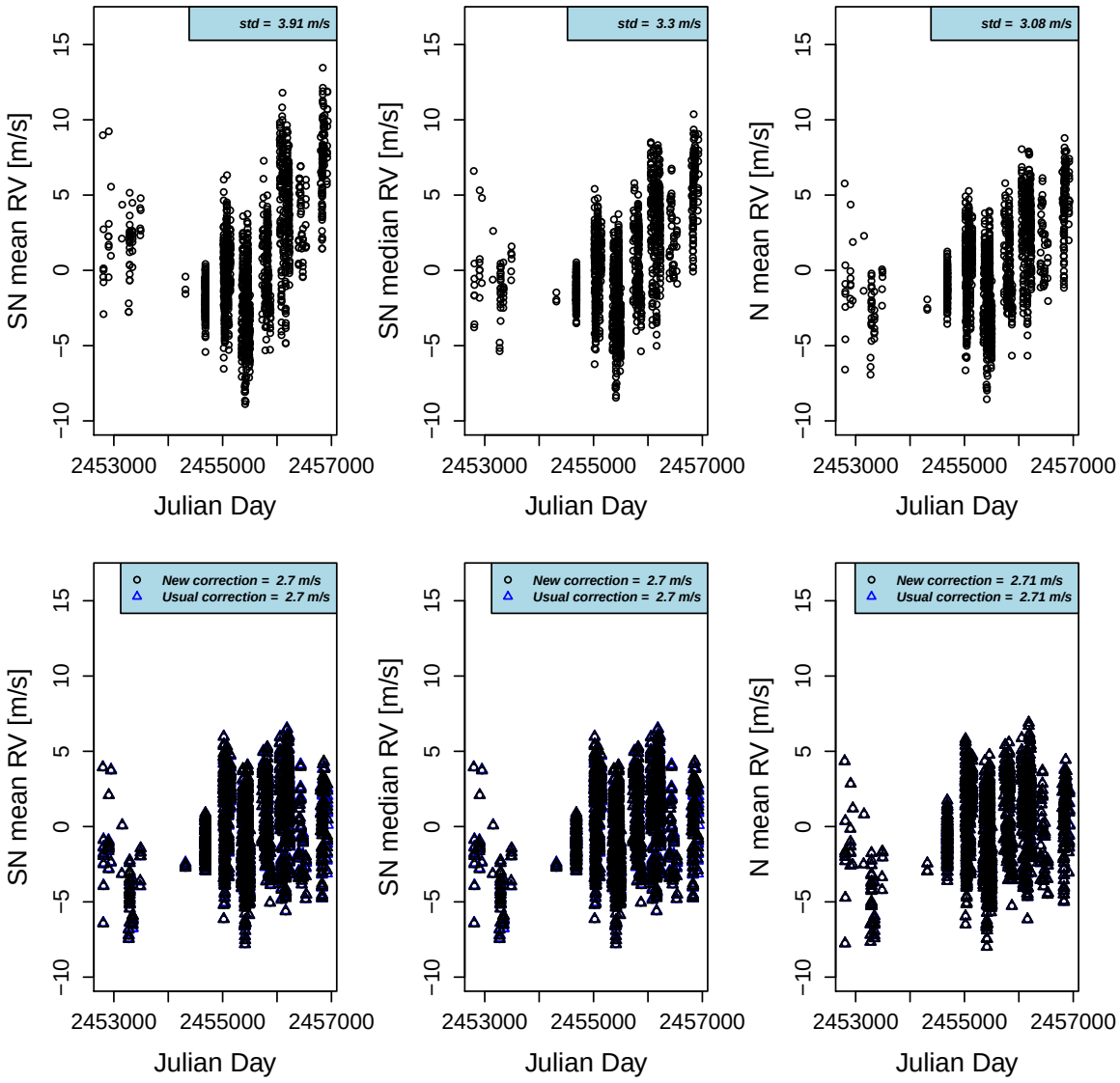


**Fig. A.2.** Set of RV's for HD192310 using a Normal and a SN fit before and once corrected from stellar activity. The correction is done using Eq. 5. Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable.

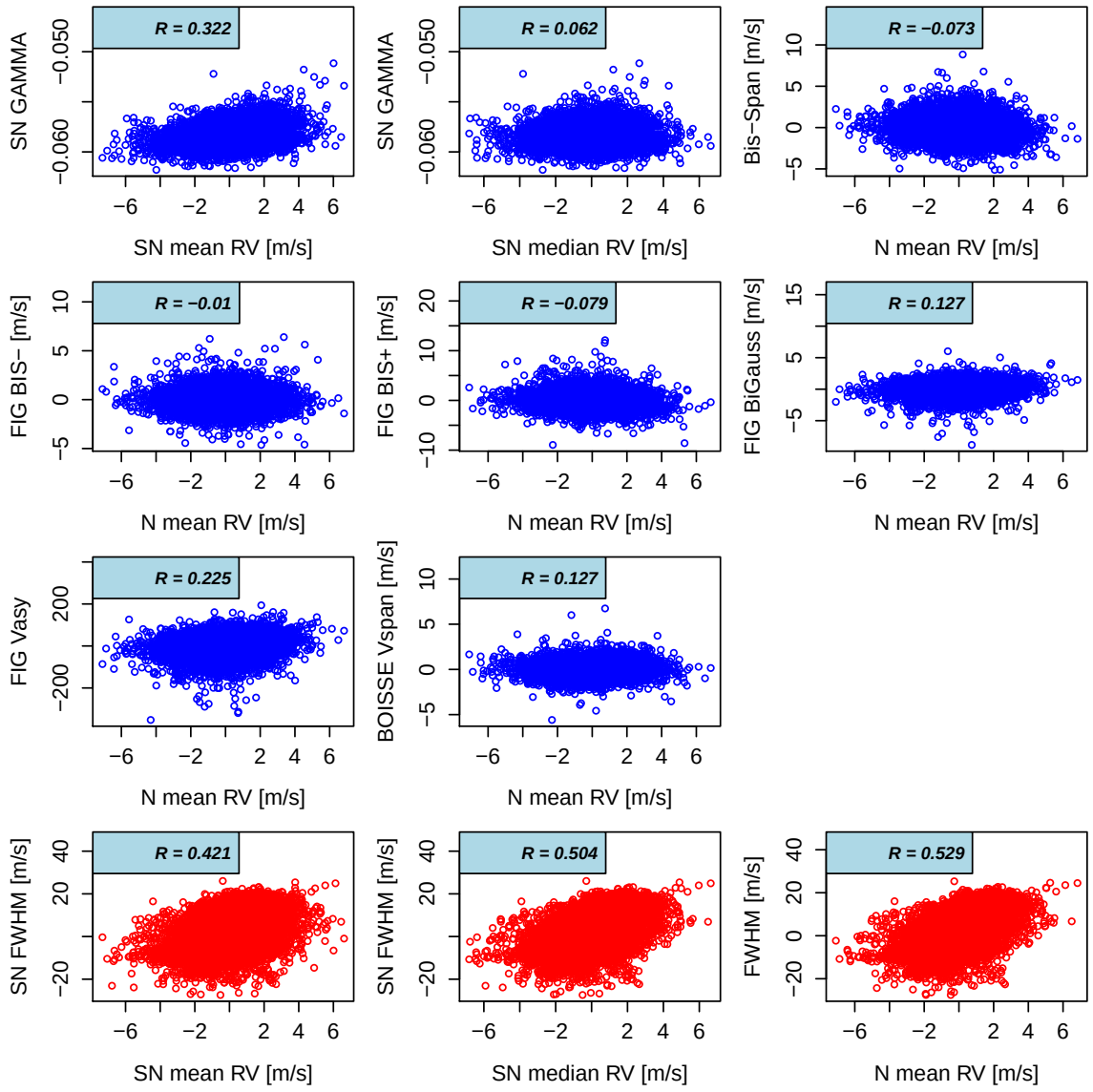


**Fig. A.3.** Correlation between the asymmetry parameters and the RV's for HD10700. The last three plots show the correlation between the FWHM and the RV's for HD10700 using respectively the SN and the Normal fits. The p–values associated with each $R$ is statistically different from 0.

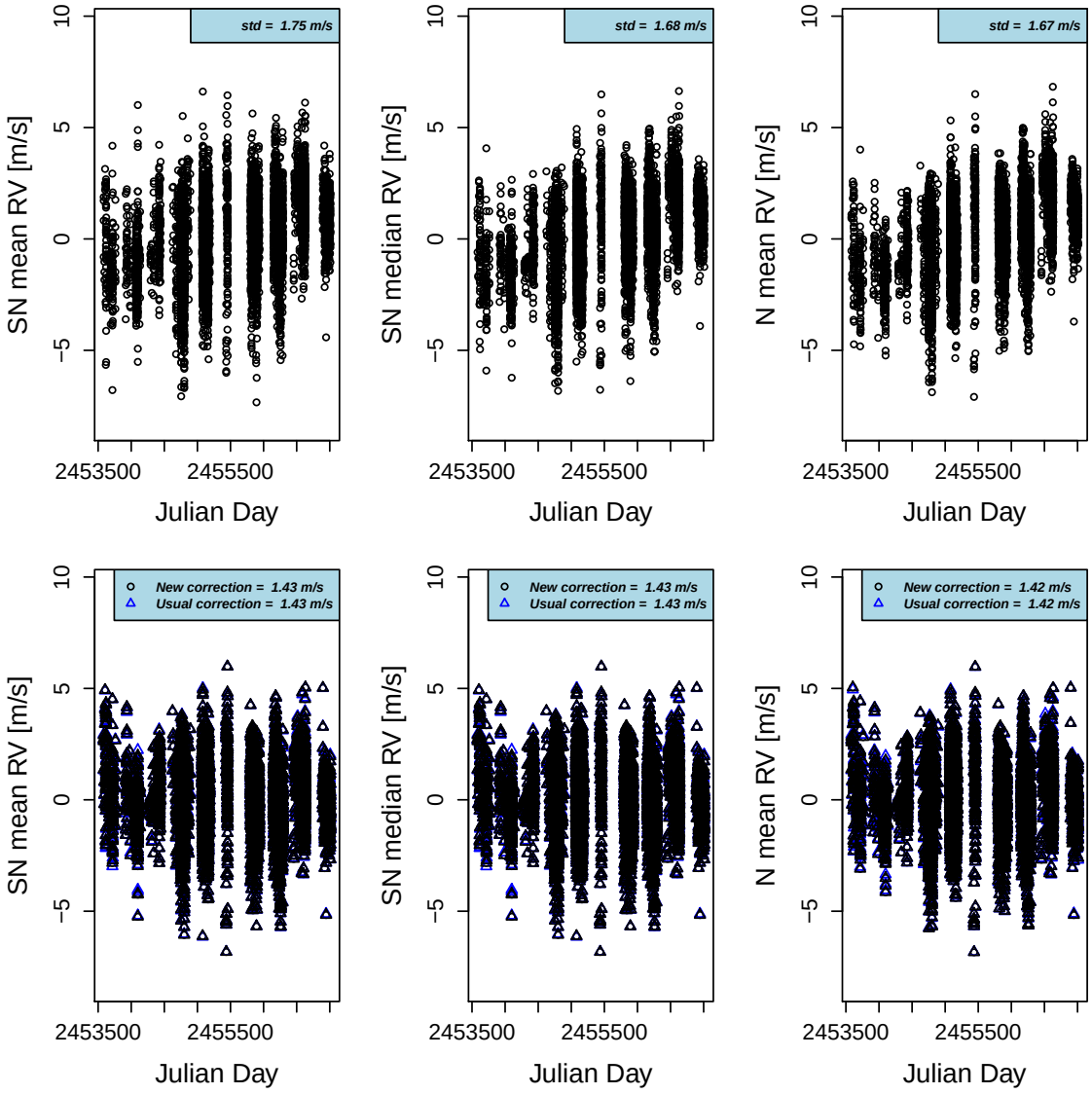


**Fig. A.4.** Set of RV's for HD10700 using a Normal and a SN fit before and once corrected from stellar activity. The correction is done using Eq. 5. Once corrected for stellar activity, the residuals in the Normal and SN analyses are comparable.

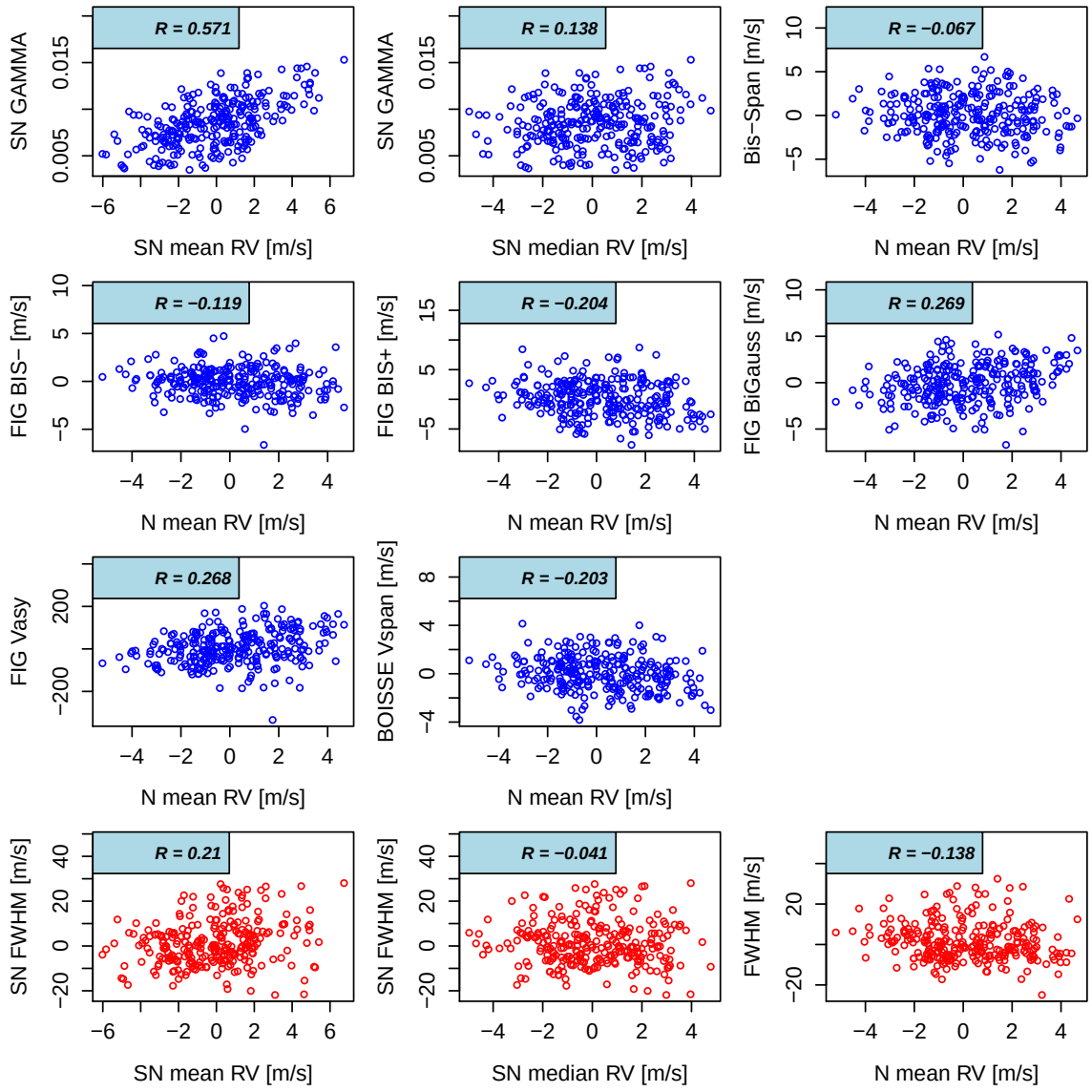


**Fig. A.5.** Correlation between the asymmetry parameters and the RV's for HD215152. The last three plots show the correlation between the FWHM and the RV's for HD215152 using respectively the SN and the Normal fits. Concerning the asymmetry of the CCF, note that the p–values associated with *R* are strongly different fro 0 for those parameters retrieved by using the SN.

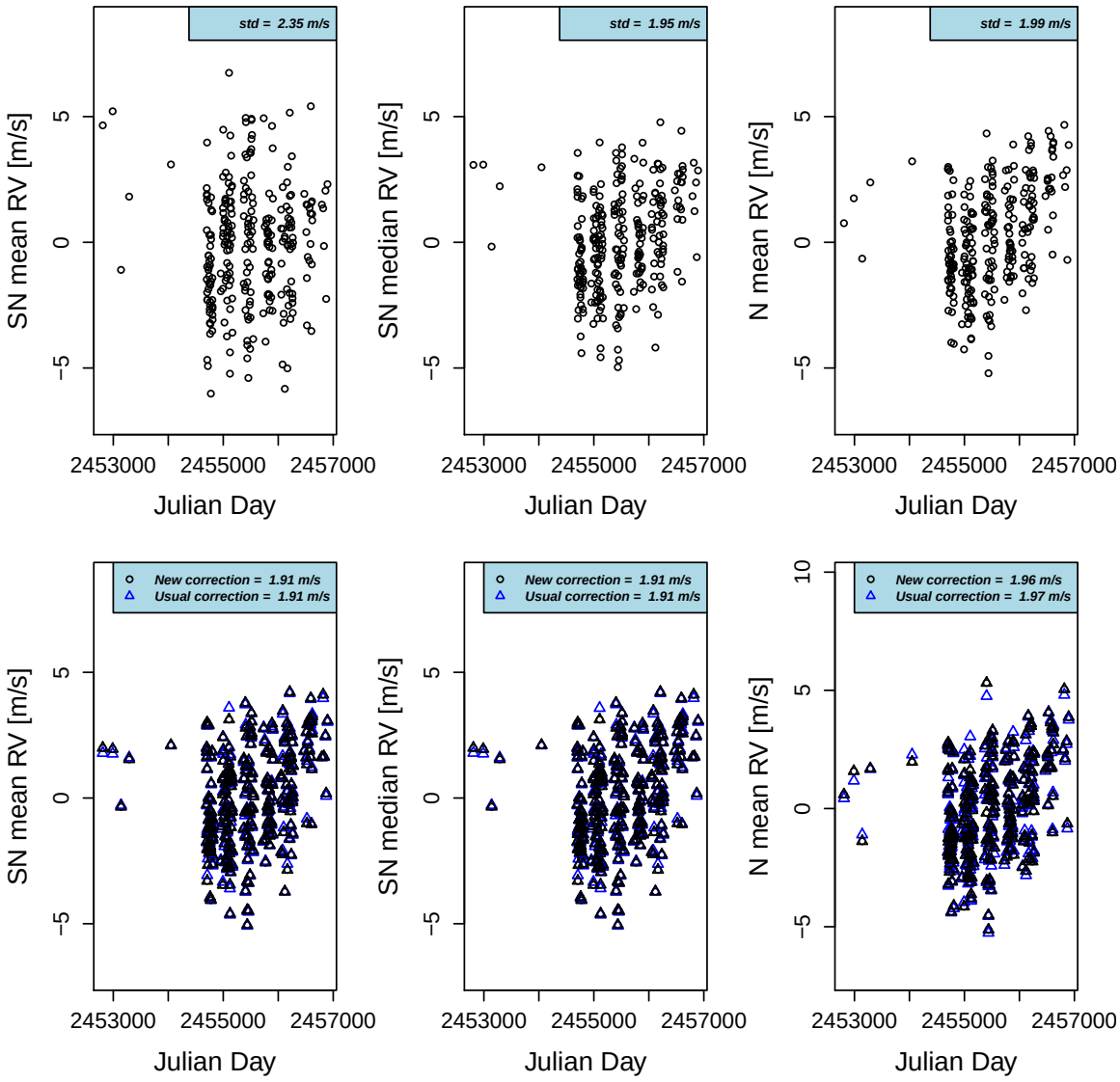


**Fig. A.6.** Set of RV's for HD215152 using a Normal and a SN fit before and once corrected from stellar activity. The correction is done using Eq. 5. Once corrected for stellar activity, the residuals for the Normal are 0.054 m s$^{-1}$higher than the residuals retrieved with the SN analysis.

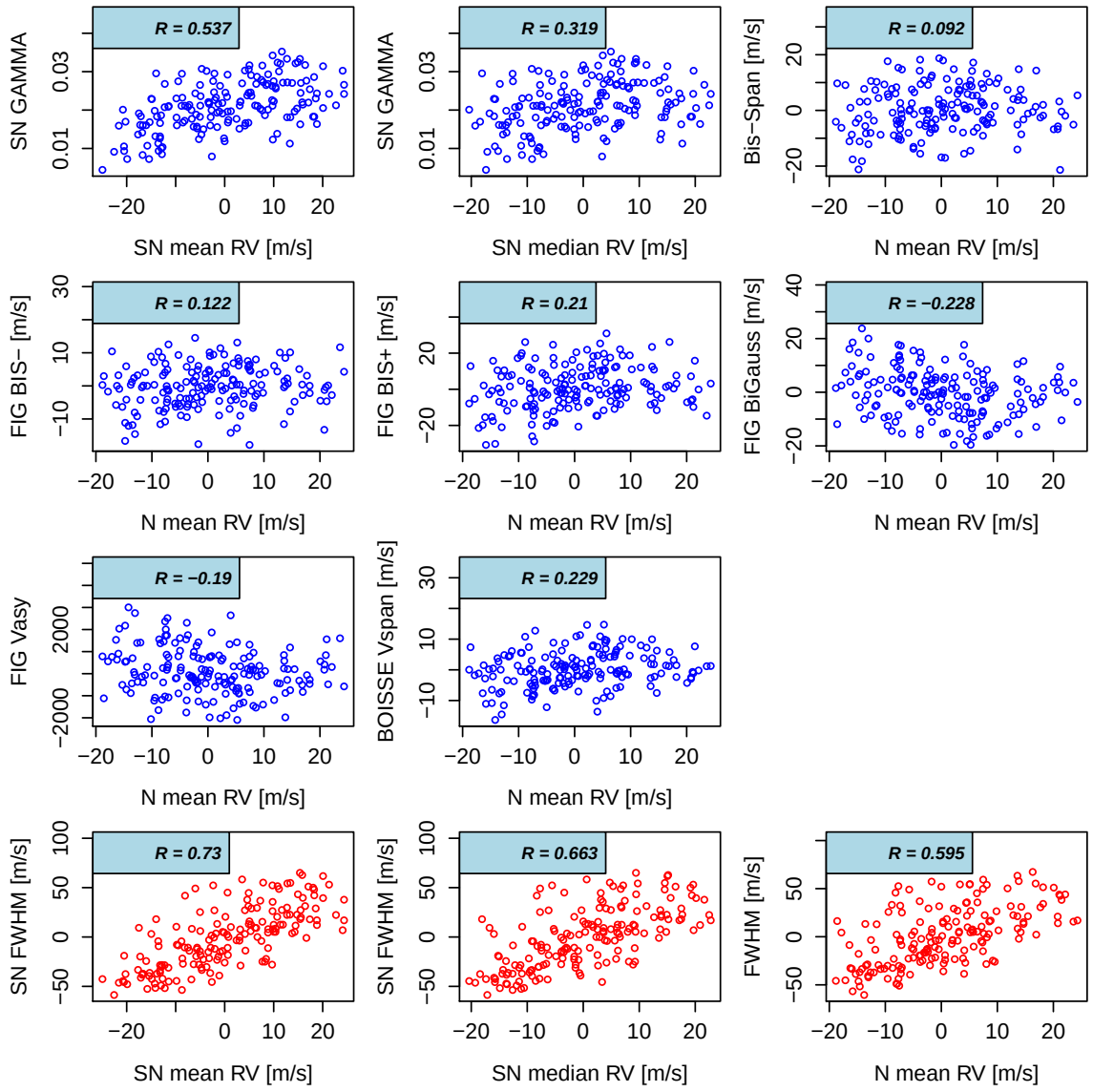


**Fig. A.7.** Correlation between the asymmetry parameters and the RV's for Corot 7. The last three plots show the correlation between the FWHM and the RV's for Corot 7 using respectively the SN and the Normal fits. Concerning the asymmetry of the CCF, note that the p–values associated with $R$ are strongly different fro 0 for those parameters retrieved by using the SN.

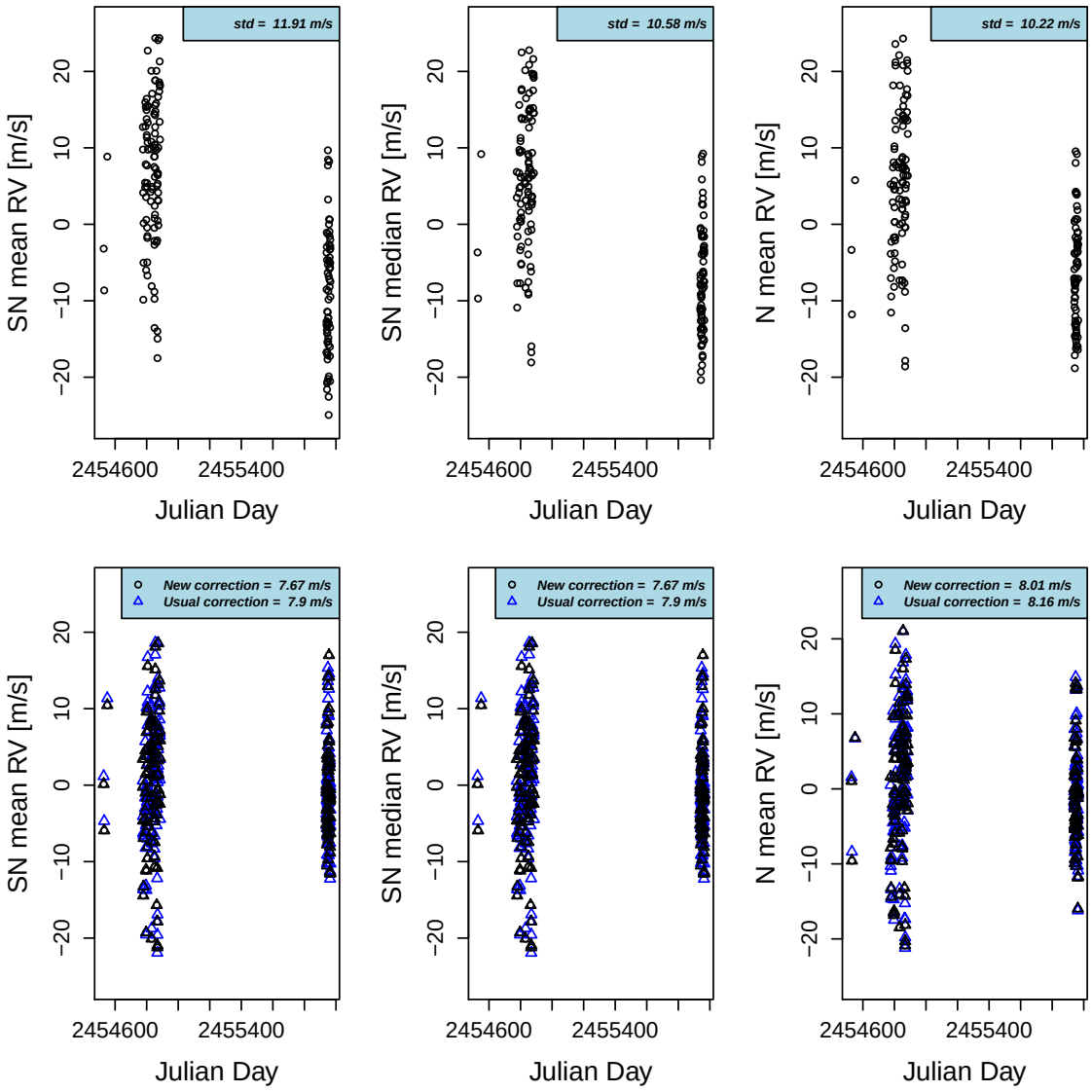


**Fig. A.8.** Set of RV's for Corot 7 using a Normal and a SN fit before and once corrected from stellar activity. The correction is done using Eq. 5. Once corrected for stellar activity, the residuals for the Normal are 0.336 $m\,s^{-1}$higher than the residuals retrieved with the SN analysis.